# The influence of speaker origin and individuality on rhythmic features of non-native speech

Thesis (cumulative thesis)
Presented to the Faculty of Arts and Social Sciences
of the University of Zurich
for the Degree of Doctor of Philosophy

by

Marie-José Kolly

Accepted in the spring semester 2016
on the recommendation of the doctoral committee:

Prof. Dr. Volker Dellwo, main supervisor
Prof. Dr. Stephan Schmid
Prof. Dr. Francis Nolan
Prof. Dr. Martin Meyer
Dr. Philippe Boula de Mareüil

Zurich, 2017

# Abstract

How do rhythmic and temporal features of non-native speech allow us to take guesses about a speaker's origin or identity? Research showed temporal phenomena to correlate with non-native speakers' intelligibility and strength of foreign accent; however, the contribution of rhythmic cues to the identification of non-native speaker origin is not yet completely understood. Likewise, little research has investigated speaker-individuality in non-native speech.

We used a perceptual approach to study the influence of speaker origin on non-native speech, and applied a range of signal manipulation methods to reduce French- and English-accented German so as to convey different types of primarily or exclusively temporal information. Findings revealed that listeners can identify speaker origin above chance in non-native speech containing temporal information alone. Our results suggest a weighting of cues in foreign accent identification, an additive trend of time and frequency domain information, perceptual salience of French-accented German and variability due to speakers.

We further investigated the influence of speaker-individuality on non-native temporal features from a speech production point of view, finding high between-speaker variability in speakers' native Zurich German and in their non-native French and English. Speaker-individual behavior was also evident within speakers, with most speakers exhibiting proportionally constant behavior in Zurich German, French, and English.

# Zusammenfassung

Wie erlauben es zeitliche und rhythmische Merkmale fremdsprachlicher Äusserungen, Sprecherherkunft oder -identität zu erkennen? Zeitliche Phänomene beeinflussen die Verständlichkeit und Akzentstärke von Nicht-Muttersprachlern, aber der Beitrag rhythmischer Information für die Identifizierung sprachlicher Herkunft ist bisher nicht geklärt. Auch hat nur wenig Forschung sprecherindividuelle fremdsprachliche Merkmale untersucht.

Mit Perzeptionsexperimenten wurde der Einfluss von Sprecherherkunft auf fremdsprachliche Äusserungen untersucht. Methoden der Signalmanipulierung reduzierten Deutsch mit französischem und englischem Akzent so, dass verschiedene Arten primär oder ausschliesslich zeitlicher Merkmale zurückblieben. Hörer konnten Sprecherherkunft aufgrund rein zeitlicher Merkmale überzufällig gut identifizieren. Weitere Ergebnisse suggerieren eine Gewichtung der Merkmale zur Akzentidentifizierung, Additivität zeitlicher und spektraler Merkmale, die spezielle Auffälligkeit von Deutsch mit französischem Akzent und sprecherspezifische Variabilität.

So untersuchte diese Arbeit auch den Einfluss von sprecherindividuellen auf zeitliche Merkmale, und fand starke Variabilität zwischen Sprechern sowohl in muttersprachlichen, zürichdeutschen, als auch in fremdsprachlichen, französischen und englischen Äusserungen. Dieses sprecherspezifische Verhalten trat insbesondere innerhalb von Sprechern hervor, indem sich diese sprachübergreifend meist proportional ähnlich verhielten.

# Acknowledgments

The main research activities presented in this PhD thesis have been conducted at the Phonetics Laboratory at the Department of Comparative Linguistics of the University of Zurich. A grant from the Swiss National Science Foundation (see below) allowed me to conduct a final experiment at the Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI) of the Centre National de la Recherche Scientifique (CNRS), in Orsay, Paris. I am truly grateful for these opportunities and I would like to thank all of those who contributed to this project.

In particular, I would like to express my sincere gratitude to:

- ◇ My main supervisor, Prof. Dr. Volker Dellwo, for giving me the chance to work on the research project *Forensic phonetic speaker identification based on temporal evidence*, and for his guidance, enthusiasm, and support during this thesis. I am indebted to him for creative ideas, for giving insightful advice, and for the many long and fruitful discussions on data.
- ◇ My co-supervisor, Prof. Dr. Stephan Schmid, for raising my interest in phonetics and phonology a long time ago, for technical advice, for precious feedback on ideas and manuscripts, and for his continual support.
- ◇ Dr. Adrian Leemann, for valuable discussions and advice regarding experiments and manuscripts as well as for his long-standing encouragement. And for carrying out the perception experiment on accent strength with English subjects.
- ◇ Dr. Philippe Boula de Mareüil, for inviting me to the LIMSI and for supervising my work in Paris, which included valuable advice on signal manipulation and numerous helpful critical discussions on methods and stimulus creation.
- ◇ Prof. Dr. Francis Nolan, for helpful discussions on the potential of pausing measures for forensic phonetics, and for accepting to be part of my thesis committee and traveling all the way from Cambridge to Zurich for this occasion.
- ◇ Prof. Dr. D. Robert Ladd, for insightful discussions on the design of the perception experiment on accent strength.
- ◇ Camilla Bernardasci and Andrea Fröhlich, for their help with some of the data collection in Zurich.

◇ Julie Belião and Ilaine Wang, for their help with recruiting subjects in Paris.

◇ Thomas Kettig, for thoroughly proofreading the English of this manuscript (any remaining mistakes are entirely mine).

◇ The members of the Phonetics Laboratory and the Department of Comparative Linguistics of the University of Zurich, as well as the members of the LIMSI, who contributed to generate an inspiring, productive, and pleasant working environment.

◇ All the subjects who took part in the experiments conducted for this thesis; this includes a high number of speakers and listeners who accepted to invest some of their time and patience.

Finally, I would like to express my deep gratitude to my parents, my sister, and my close friends for their long-standing support during this period, sharing many of the joyful moments and successes as well as all of the more difficult times with me.

This research was funded by two grants from the Swiss National Science Foundation (SNSF): The project grant number 100015_135287 attributed to Prof. Dr. Volker Dellwo and Prof. Dr. Stephan Schmid, and the Doc.Mobility grant number P1ZHP1_155024 attributed to myself. Some of these research activities were additionally possible thanks to a grant from the Department of Comparative Linguistics of the University of Zurich. I am very grateful to these two institutions for their trust and support.

Zurich, January 2016
Marie-José Kolly

# Contents


**I General introduction** **1**

**1 Motivation** **3**
1.1 Aims of the thesis . . . . . 5
1.2 Outline of the thesis . . . . . 6

**2 Research on temporal and rhythmic features of speech** **9**
2.1 Research on speech rhythm . . . . . 10
2.1.1 Research on native speech rhythm . . . . . 10
2.1.2 Research on non-native speech rhythm . . . . . 11
2.1.3 Research on speaker-individual speech rhythm . . . . . 13
2.2 Speech corpora used for the present work . . . . . 15

**II Experimental investigations** **17**

**3 Temporal and rhythmic features in German as spoken by French, English and Zurich German speakers** **19**

**4 Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition** **27**

**5 Foreign accent recognition based on temporal information contained in lowpass-filtered speech** **41**

**6 Listeners use temporal information to identify French- and English-accented speech** **49**

**7 Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study** **65**

**8 Strength of foreign accent is speaker-specific across different non-native languages** **73**

**III General discussion and conclusions 81**

**9 General discussion 83**
9.1 Cues to speaker origin in the time domain . . . . . 83
9.2 Evidence for speaker-individuality in the time domain . . . . . 89
9.3 General properties of non-native speech in the time domain . . . . . 91
9.4 Possible implications of the present work . . . . . 92

**10 Conclusions and outlook 95**
10.1 Main contributions . . . . . 95
10.2 Future work . . . . . 96
10.3 Concluding remarks . . . . . 97

**A Appendix: Reading materials 99**

**B References 103**

# Part I

# General introduction

# 1

# Motivation

"*Judging by your accent, you must be French.*" Linguistic origin is a frequently discussed topic in everyday social interactions. A few syllables or even sounds are often sufficient for listeners to detect, and possibly identify, a foreign accent. Among other characteristics specific to foreign accents, non-native speech has been claimed to differ from native speech in its rhythmic and temporal structure, influenced by temporal features of a speaker's native language (Lloyd James, 1929) as well as individually unique durational features (Derwing et al., 2009). As such, three groups of features could be said to influence non-native speech, causing it to differ from native speech:

(i) *Interference from a speaker's native language.* A typical example is voice onset time, the native-like acquisition of which has been observed to be a difficult task for learners (Caramazza et al., 1973; Flege and Hillenbrand, 1984; Hazan and Boulakia, 1993).

(ii) *Speaker-individuality.* Two non-native speakers with comparable linguistic biographies and non-native proficiency might have different habits, preferences, or constraints regarding the realization of particular acoustic patterns. Very little research has focused on speaker-idiosyncratic features of non-native speech. Studies have found non-native fluency to be speaker-individual to some extent, where non-native fluency exhibits features similar to individuals' native fluency (Derwing et al., 2009; de Jong et al., 2013).

(iii) *General properties of non-native speech independent from a specific native language or speaker.* For example, it has been observed that non-native speakers often produce more vocalic material than native speakers (White and Mattys, 2007a; Dellwo, 2010), possibly because they tend to lengthen the duration of vowels (Adams and Munro, 1978; Taylor, 1981; Derwing et al., 2009).

This thesis addresses rhythmic and temporal features of non-native speech with regard to features from groups (i) and (ii), though the work conducted in this context further allows the description of certain features from group (iii).

Are listeners able to identify speaker origin based on speech rhythmic and temporal features of non-native speech? As to features from group (i), research on non-native speech production suggests that non-native speakers produce rhythmic features in an "intermediate" fashion, between their native language and the target language. However, there is a growing body of evidence that challenges this (see Section 2.1.2). In addition, the measures used to quantify rhythm in most of these studies have been shown to be heavily influenced by a number of factors other than the language variety spoken, with variability in such measures sometimes as high within a language as between languages (see Section 2.1.1). Furthermore, it has been argued that speech rhythm may be a primarily perceptual phenomenon (Lloyd James, 1929; Lehiste, 1977; White et al., 2012). This thesis therefore adopts a perceptual approach to the investigation of speaker origin and its influence on non-native rhythmic and temporal features. Languages are widely acknowledged to differ in their rhythmic or (suprasegmental) temporal organization, and certain durational patterns of non-native speech have been shown to arise from interference from a speaker's native language. Some of these features may be perceptually salient to listeners in terms of speaker origin. One motivation for the investigation of listeners' perception of speaker origin based on temporal information comes from the domain of forensic phonetics. Forensic recordings used for casework are often strongly degraded in the frequency domain, as they are typically obtained over a telephone network or distorted by background noise (Hirson et al., 1995). Because the identification of speaker origin is often a significant element of casework, time domain features may be able to complement analyses and contribute to conclusions about a speaker's profile (Ellis, 1994; Köster et al., 2012). In particular, it may be useful to better understand how reduced frequency domain information affects listeners' accent identification performance, and whether certain foreign accents are more readily identifiable based on temporal and rhythmic features than others.

Do speech rhythmic and temporal features reveal speaker-individuality across different languages? As to features from group (ii), research on native speech has shown that speech temporal and rhythmic features vary considerably between speakers from the same linguistic background. A number of such features reveal both high between-speaker variability and relative robustness within speakers, even when speakers are investigated using different speaking styles, disguising their dialect, speaking in different intended tempos, and recording their voices through different channels. Furthermore, some features remain speaker-specific when balanced bilinguals produce speech in different languages. Moreover, speaker-individual features in the native language have been shown to influence temporal measures of non-native fluency (see Section 2.1.3). It is therefore conceivable that some temporal and rhythmic patterns vary between speakers but remain in part unchanged whether a speaker is speaking a native or non-native language. Such phenomena could be leveraged for forensic phonetic purposes: in forensic voice comparison, cases occur where there is a mismatch between the language of the acoustic trace and that of the comparison material. However, the impact of speaking a non-native language on speaker-individual features is largely unknown, and forensic phoneticians are advised to "exercise particular caution" in such cases (International Association for Forensic Phonetics and Acoustics, 2004). Because some perpetrators or suspects use a non-native language as

a means to disguise their voice and because frequency domain characteristics are often degraded in casework material, information on temporal features of native and non-native speech may complement the already existing speaker-specific parameters used in forensic phonetics (Dellwo et al., 2012; Leemann et al., 2014).

Do non-native speakers differ from native speakers in their temporal and rhythmic features, regardless of their native vs. non-native language combination? As to features from group (iii), research has pointed to a number of temporal measures, particularly measures of fluency, that seem to be typical of non-native speech. The experiments conducted in the context of this thesis confirm a number of results previously found in this domain and point to some additional temporal features that may be characteristic of non-nativeness.

## 1.1 Aims of the thesis

In order to contribute to the characterization of non-native speech in the temporal and rhythmic domain, four main objectives have been defined for this thesis.

**Aim 1: Development of experimental setup and signal manipulation methods suited to presenting listeners with speech temporal features.** The experimental setup for such perception experiments should involve an appropriate choice of speech material, that is: (i) non-native accents that can be assumed to differ in their temporal or rhythmic structure; (ii) sufficient variability in speaker and sentence material such that conclusions can be generalized to a larger population of speakers and linguistic material; and (iii) similar accent strength between the speakers of the different non-native accents. The signal manipulation methods adopted and developed for these experiments must present various types of predominantly and exclusively temporal cues within the speech material.

**Aim 2: Description of speech temporal cues that contribute to the identification of speaker origin in non-native speech.** Speech temporal features are highly multidimensional, which is why the perception experiments need to involve signal conditions that expose listeners to different types of temporal features. Furthermore, the combined presence of speech temporal and spectral features may induce additive effects on listeners' accent recognition performance. Listeners' sensitivity to temporal cues should therefore be investigated when spectral information is reduced to various degrees, and when only one type of cue — temporal or spectral — is present in speech material. Furthermore, it is important to study the effect of particular foreign accents' as well as of particular speakers' temporal patterns on listeners' accent identification performance. The perception experiments should be complemented by a description of the temporal patterns contained in the non-native speech material from a speech production point of view.

**Aim 3: Development and description of a speech corpus suited for the investigation of speaker-individual, language-invariant temporal patterns.** Such a corpus should involve a number of languages and represent speakers' native as well as non-native speech. Since the corpus is designed to study speaker-individual features, it should contain speech from a relatively high number of speakers who should be homogenous regarding a number of variables, such that potential between-speaker differences can in fact be attributed to speaker-individuality. Therefore, speakers should be compared within groups sharing a common background regarding not only their native, but also their non-native languages, age, education, and (balanced for) gender. The corpus should further contain a reasonable amount of speech material for each language and speaker such that potential effects cannot be attributed to the phonetic or phonological features of specific sentences. Little research has investigated language-invariant speaker-specific features, and temporal patterns are known to vary with the specific linguistic material involved, so this corpus needs to be constructed with highly controlled contexts. Therefore, all speakers should, as much as possible, produce the same segmental material. This is why read speech is most suited for this corpus.

**Aim 4: Description of temporal patterns that remain speaker-individual in a cross-linguistic context.** Again, as temporal cues are highly multidimensional, a wide range of durational features could potentially be investigated between speakers and across languages. The present thesis investigates features related to pausing behavior, measures of speaker fluency that have been suggested to exhibit speaker-individuality in non-native speech. As such measures typically vary with the linguistic material involved, the effect of sentence as well as the effects of language and speaker need to be investigated. Also investigated here is speakers' foreign accent strength as rated by native listeners, allowing for between- and within-speaker investigations of accent strength. The application of further durational measures to this corpus is planned for the future.

## 1.2 Outline of the thesis

Part I first motivates the choice of research questions in Chapter 1. Chapter 2 includes a general introduction on speech rhythmic and temporal features and a presentation of the speech material used to investigate the questions outlined above. In particular, Section 2.1.1 gives a brief overview of the literature on speech rhythm, and research on non-native speech rhythm and speaker-individual temporal features are reviewed in Sections 2.1.2 and 2.1.3.

Part II comprises four research papers and two smaller experiments. The research papers present a series of perception experiments that investigate the influence of speaker origin on temporal patterns of non-native speech, and a production experiment that investigates the influence of speaker-individuality on durational features in foreign accented speech. While the research papers cover the influence of speaker origin from a perceptual perspective, the experiment presented in Chapter 3 characterizes the temporal patterns of

the speech material from the point of view of production. The perception experiments presented in Chapters 4, 5, and 6 investigate whether listeners can identify speaker origin in non-native speech based on primarily or exclusively temporal characteristics. In Chapter 4 we present results on the influence of various types of temporal features as well as of the amount of frequency domain information present in stimuli on listeners' accent identification performance. The experiment reported in Chapter 5 takes a somewhat different approach by presenting listeners with temporal and rhythmic information in signal types specifically designed to sound familiar to listeners. Chapter 6 presents an experiment that takes this idea one step further, designing stimuli so as to sound as close to natural speech as possible. This experiment investigates the additive effect of temporal and spectral features on listeners' accent recognition performance by presenting listeners with each type of cue separately. The production experiment in Chapter 7 examines the influence of speaker-individuality on measures of pausing in a cross-linguistic study involving speakers' native language as well as two of their second languages. Finally, Chapter 8 presents a perception experiment conducted to elicit ratings on speakers' accent strength as well as results on speaker-individuality in this variable.

Part III of this thesis includes a general discussion of all the experiments presented in Part II. It further summarizes the outcome of this thesis with respect to its aims, highlights the novel contribution of the present work, and provides an outlook on possible future investigations.

# 2

# Research on temporal and rhythmic features of speech

Amongst many other definitions, speech can be thought of as acoustic energy that is distributed over a number of specific frequency ranges; this distribution changes over time, governed to a high degree by language-, dialect-, accent- and speaker-specific characteristics. As exemplified in the spectrogram presented in Figure 2.1, the frequency domain of an acoustic signal encodes information on the intensity of the acoustic energy in a given range of frequencies, i.e., spectral information. The time domain represents how the signal, i.e., the energy present in different frequency ranges, changes over time.

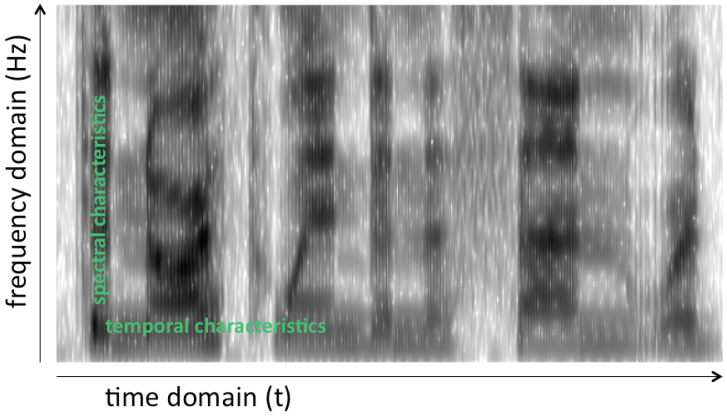



Figure 2.1: Spectrogram illustrating the time domain (x-axis) and the frequency domain (y-axis) of an acoustic signal.

The frequency domain of speech contains information about voice quality, fundamental frequency, and vowel and consonant qualities, for example. The time domain, on the other hand, contains information about the durations of segments, syllables, intonation phrases, and any other element represented in the frequency domain. Compared to research on frequency domain characteristics, the time domain of speech has gained less attention in phonetic research. Though there is no consensus to how rhythm is encoded in the acoustic

signal, speech rhythmic characteristics have typically been attributed to the time domain. It is unclear, however, whether rhythm is created by the periodic recurrence of one particular element or pattern (*coordinative* rhythm), by the alternation of strong and weak elements (*contrastive* rhythm), or whether it is a metaphor for patterns that are neither 'periodic' nor 'alternating' in a strict sense (Nolan and Jeon, 2014). Furthermore, the perceptual impression of rhythmicity has been shown to be multidimensional, with duration, intensity, fundamental frequency, and even spectral information playing a possible role both in the perception of temporally more or less regular patterns and in the perception of the alternation of more or less prominent elements in the speech signal (Cummins and Port, 1998; Lee and Todd, 2004; Dellwo, 2008; Tilsen and Johnson, 2008; Barry et al., 2009; Kohler, 2009; Tilsen and Arvaniti, 2013).

The experiments presented in this thesis explore time domain characteristics to learn about (i) the ways in which these may encode speaker-individuality across different languages, and (ii) whether they may, possibly in combination with frequency or intensity domain characteristics, be useful for the identification of foreign accents. The time domain phenomena investigated contribute to a perceptual impression of rhythmicity to some extent, but we do not make the claim that these phenomena create a comprehensive impression of rhythmicity on their own. The following sections present a brief general review of research on speech rhythm in native speech (Section 2.1.1), on the influence of non-nativeness on speech rhythmic features (Section 2.1.2) and on the influence of speaker-individuality on speech temporal features (Section 2.1.3). More specific literature reviews considering each of the conducted experiments can be found in Section 1 of Chapters 4, 5, 6, and 7.

## 2.1 Research on speech rhythm

### 2.1.1 Research on native speech rhythm

The work of Lloyd James (1929), Pike (1945), and Abercrombie (1967) has established that languages differ in their suprasegmental temporal or rhythmic organization. The so-called Rhythm Class Hypothesis, which suggests that some languages are characterized by equal syllable durations (e.g., French, Spanish, Italian) and others by equal distances between stressed syllables (e.g., German, English, Dutch), is heavily disputed, however (Dellwo, 2010; Loukina et al., 2011; Arvaniti, 2012). The acoustic correlates for rhythmic classes — interstress intervals and syllables — were initially claimed to be isochronic, i.e., of equal length. After the isochrony hypothesis was empirically disproved (Roach, 1982; Dauer, 1983), such acoustic correlates of perceived rhythm were sought in the durational variability of consonantal and vocalic intervals, which was assumed to reflect phonological differences between languages — typically, different degrees of syllable complexity and vowel reduction (Dasher and Bollinger, 1982; Dauer, 1983). Based on these acoustical units, so-called rhythm metrics were constructed to capture speech rhythm and to illustrate differences between languages of different rhythmic classes in the time domain

(Ramus et al., 1999; Low et al., 2000; Grabe and Low, 2002; Dellwo, 2006; Dellwo et al., 2007; White and Mattys, 2007a; see Loukina et al., 2011 and Chapter 3 for an overview).

However, consonantal and vocalic intervals do not constitute the only acoustical unit relevant for temporal variability in speech; information structured in the time domain is highly multidimensional. Therefore, different acoustic characteristics have been used for the construction of rhythm metrics. For example, Dellwo et al. (2007) presented an alternative acoustical unit by applying rhythm metrics to voiced and unvoiced intervals instead of vocalic and consonantal intervals. Asu and Nolan (2006) and Nolan and Asu (2009) have applied rhythm metrics to syllables and feet, in reference to the claims made by the initial Rhythm Class Hypothesis as to the isochrony of syllables and interstress intervals. Recent attempts have also included information from the intensity domain (Low, 1998; Dellwo et al., 2012; Tilsen and Arvaniti, 2013) and fundamental frequency information (Cumming, 2011), since it has been argued that perceived speech rhythm not only emerges from durational characteristics, but also from intensity, fundamental frequency, and spectral dynamics (Kohler, 2009).

Furthermore, it has been shown that rhythm metrics capture not only durational differences between languages (Ramus et al., 1999; Grabe and Low, 2002; Dellwo, 2006; Dellwo et al., 2007) but also between dialects (varieties of English: Low et al., 2000; Ferragne and Pellegrino, 2004; White and Mattys, 2007b; Irish dialects: Dorn et al., 2012; Swiss German dialects: Leemann et al., 2012; Italian dialects: Schmid, 2012; French dialects: Obin et al., 2012), individual speakers (Dellwo, 2010; Wiget et al., 2010; Yoon, 2010; Loukina et al., 2011; Arvaniti, 2012; Dellwo et al., 2012, 2015; Leemann et al., 2014) and sentence material (Dellwo, 2010; Wiget et al., 2010; Arvaniti, 2012). This has raised questions as to whether or not such metrics are suited for characterizing languages as belonging to a particular rhythmic class (Arvaniti, 2012), and as to whether such rhythmic classes even exist at all (e.g., White and Mattys, 2007b). Though the investigation of rhythm in speech production may be hotly debated, studies of speech perception consistently attest to perceivable differences between languages (Nazzi et al., 1998; Ramus and Mehler, 1999; Ramus, 2002; Ramus et al., 2003) and dialects (White et al., 2012). In such experiments, listeners are not necessarily sensitive to rhythmic classes but rather to specific suprasegmental and segmental durational cues (White et al., 2012). As suggested by Classe (1939) and Lehiste (1977), the heavily discussed isochrony is probably primarily a perceptual phenomenon, in which languages seem to be perceived as being more (e.g., French) or less (e.g., English, German) regularly timed (Dellwo, 2008).

### 2.1.2 Research on non-native speech rhythm

Perception studies have revealed that non-native durational patterns decrease speakers' intelligibility and increase the perceptual impression of foreign accent strength (Tajima et al., 1997; Munro and Derwing, 2001; Bent et al., 2008; Holm, 2008; Dellwo, 2010; Pinet and Iverson, 2010; Quené and van Delft, 2010; Winters and O'Brien, 2013). The idea that speech temporal and rhythmic features play a role in the perception of non-native

speech has a long tradition, dating to Lloyd James' (1929) suggestion that non-native speech rhythm strongly affects the intelligibility of foreign-accented speech.

Lloyd James (1929) argued that non-native speech is characterized by rhythmic interference from a speaker's native language and that such interference affects perception more strongly when languages differ in their rhythmical organization. Later research has shown that, indeed, non-native durational patterns do not only reflect non-nativeness in general, but are influenced by the temporal patterns of a speaker's specific native language. Such interference has been found for durational phenomena such as the production of vowel quantity (McAllister et al., 2002), voice onset time (Caramazza et al., 1973; Flege and Hillenbrand, 1984; Hazan and Boulakia, 1993; Fowler et al., 2008), and word-final stop closure duration (Flege et al., 1992; Arslan and Hansen, 1997). These segmental or subsegmental temporal features of non-native speech are likely to translate into suprasegmental temporal features, as it has been shown that durations of suprasegmental units depend on intrinsic durations of the segments they contain (van Santen and Shih, 2000). Such suprasegmental temporal features could in turn be assumed to influence the duration-based rhythm metrics mentioned in Section 2.1.1, as these metrics capture variability in a range of suprasegmental durational features. Assuming (a) that rhythm metrics do in fact capture rhythm, (b) that non-native speech rhythm is characterized by interference from a speaker's native language, and (c) that some languages are more similar in their rhythmical organization than others (e.g., French, Spanish, Italian vs. German, English, Dutch), one would expect the following: non-native speakers' values for rhythm metrics should fall in between values attested for the native and the target language for language pairs with different rhythmical organization, and these values should be similar to both the native and the target language for language pairs with similar rhythmical organization.

Research has revealed that this is the case for some rhythm metrics (but not for others) and for some native vs. non-native language pairs (but not for others). For example, values for the rate-normalized durational variability of vocalic intervals (*varcoV*, White and Mattys, 2007a) measured in English speakers of Spanish and Spanish speakers of English fall in between the values found for native Spanish and native English (Carter, 2005; White and Mattys, 2007a; Gutiérrez Díez et al., 2008). The same holds for the rate-normalized durational variability of adjacent vocalic intervals in Spanish speakers of Afrikaans and Afrikaans speakers of Spanish (*nPVI_V*; Coetzee et al., 2015). English and Dutch, however, exhibit very similar values for native and target language as well as for non-native speech (White and Mattys, 2007a). This is in line with Spanish learners of English being perceived to have a stronger accent than Dutch learners of English (White and Mattys, 2007a). These findings are in line with expectations, as English, Dutch and Afrikaans are assumed to differ from Spanish in their rhythmical organization (see Section 2.1.1). Furthermore, data from Gutiérrez Díez et al. (2008) for Spanish learners of English, from Tortel and Hirst (2010) for French learners of English, and from Ordin and Polyanskaya (2015) for French and German learners of English has shown that while learners' values for certain rhythm metrics fall between the native and the target language, more advanced learners exhibit more target-like values.

Other findings, however, challenge these ideas. English learners of Spanish (White and Mattys, 2007a) and German learners of French and English (Dellwo, 2010) exhibit higher values than both their native and their target languages for the proportion of speech that is vocalic ($\%V$, Ramus et al., 1999). This overshoot by non-native speakers of different language backgrounds suggests that a high $\%V$ may be a general property of non-native speech. This may be explained by the finding that non-native speakers tend to lengthen the duration of vowels, particularly of unstressed vowels, giving the auditory impression of more regular speech timing (Adams and Munro, 1978; Taylor, 1981; Derwing et al., 2009). Conflicting evidence from Tortel and Hirst (2010), however, shows that French learners of English exhibit lower values for $\%V$ than their native and target language. Finally, in a study by Grenon and White (2008), Japanese learners of English and English learners of Japanese exhibited target-like values for several vocalic metrics, which is unexpected, as Japanese and English have previously been shown to differ in their rhythmic organization (Ramus et al., 1999).

Therefore, non-native speech seems to be influenced by durational characteristics of speakers' native language for some rhythm metrics and language pairs; other metrics and language pairs, however, seem to reflect general properties of non-native speech rather than specific interference from the native language. Furthermore, some of the findings reported above do not reveal readily interpretable interference from a speaker's native language, nor can they be explained by well-known properties of non-native speech. Such results are possibly due to relatively small datasets in terms of speakers and sentences; as discussed in Section 2.1.1, rhythm metrics are strongly influenced by these factors. It is also possible, however, that these metrics are not suited for application as acoustic correlates of perceived rhythmicity at all (Arvaniti, 2012). It therefore remains unclear which of the durational characteristics of non-native speech are due to interference from a speaker's native language, which are general features of non-native speech, and which can, as of yet, not be explained. It further remains widely unclear whether such acoustic temporal variability between different non-native accents is perceptually salient. While perceptually salient rhythmic differences between some languages have been empirically attested by various studies (see Section 2.1.1), the idea that such characteristics also play a role in non-native speech has been investigated empirically only for speech production. This is why Chapters 4, 5, and 6 take a speech perception point of view for the investigation of foreign-accent-specific temporal features.

### 2.1.3 Research on speaker-individual speech rhythm

Many of the metrics proposed to measure rhythm are determined not only by the language spoken, but also by the individual speaker (Dellwo, 2010; Wiget et al., 2010; Yoon, 2010; Arvaniti, 2012; Dellwo et al., 2012, 2015; Leemann et al., 2014). Whether this implies the existence of perceived speaker-specific rhythmicity is still unclear; however, there is growing evidence for speaker-individual behavior in the time domain. Apart from between-speaker variability in rhythm metrics, it is known that native speakers differ in

pausing behavior (Goldman Eisler, 1968; Künzel, 2013), in durational characteristics of their hesitation markers (Braun and Rosin, 2015), in voice onset time (Allen et al., 2003) and in the durations of syllable nuclei (Shriberg et al., 2005). It has further been shown that a number of temporal features related to fluency vary between non-native speakers. Some of this variability is correlated with speakers' fluency in their native language, a phenomenon attributed to between-speaker differences in cognitive processing (Derwing et al., 2009; de Jong et al., 2013).

Unlike frequency domain characteristics, which are known to be influenced to some extent by the anatomical characteristics of a speaker's larynx and vocal tract (Fant, 1960), the reasons for time domain characteristics to vary between speakers are not as readily interpretable. Cognitive factors, as mentioned above, have been assumed to influence between-speaker variation in fluency measures. One interpretation that may apply to other types of temporal patterns involves a comparison to the human gait, which seems to be highly idiosyncratic due to individual ways of moving one's body parts (Loula et al., 2005); McDougall (2006), Dellwo et al. (2012), and Dellwo et al. (2015) suggest that movements of the articulators may be idiosyncratic in similar ways. Since there are intrinsic — mechanical — properties of the articulators (e.g., volume, mass, velocity, shape; McDougall, 2006; Perrier, 2012), there may be speaker-individual ways of moving these articulators. Such phenomena may leave a "stamp" in the acoustic speech signal, particularly in its time domain (Dellwo et al., 2015). According to this rationale, certain temporal patterns should not only vary between speakers, but they should also be robust within a speaker to some extent, which is a prerequisite for acoustic features to be used in forensic casework (Nolan, 2009).

Experimental investigations carried out in a forensic phonetic context have revealed a number of temporal patterns that meet these requirements. McDougall (2004, 2006) has found the dynamics of formant trajectories to vary between speakers; Dellwo and Schmid (2015) find articulation rate to vary between speakers, and Dellwo et al. (2012, 2015), Leemann et al. (2014), Dellwo and Schmid (2015), and Leemann and Kolly (2015) find a number of rhythm metrics to vary between speakers. Some of these rhythm metrics are relatively invariant to within-speaker variation introduced by articulation rate (five intended tempo conditions; Dellwo et al., 2015), linguistic structural characteristics (sentences generated by the same speaker vs. by other speakers; Dellwo et al., 2015), speaking style (read vs. spontaneous speech; Leemann et al., 2014), dialect spoken (Leemann and Kolly, 2015), language spoken (bilingual speakers in their German vs. Italian speech; Dellwo and Schmid, 2015), and channel (hifi vs. telephone speech; Leemann et al., 2014). Given these findings and the anatomical rationale for them, this thesis hypothesizes that speaker-individual temporal patterns may be found even when within-speaker variation is introduced by speakers producing native and non-native speech. This idea is explored in Chapters 7 and 8 and developed further in Section 10.2.

## 2.2 Speech corpora used for the present work

Two different speech corpora came into play for the study of rhythmic and temporal features specific to speaker origin and to speaker-individuality:

(i) to investigate the perceptual identification of speaker origin based on rhythmic and temporal characteristics of non-native speech, we used Standard German speech read by French, English and Zurich German speakers — the Non-native Speaker Origin Corpus;

(ii) to investigate speaker-individual influences on rhythmic and temporal features in native and non-native speech, we used a subset of the TEVOID corpus, which contains Zurich German, French and English speech read by Zurich German speakers (Dellwo et al., 2012, 2015; Leemann et al., 2014) — the Non-native Speaker-Individuality Corpus.

Both corpora were further used to describe a number of general time domain properties of non-native speech. The two corpora are briefly introduced here; they are described in more detail in Part II of this thesis.

The Non-native Speaker Origin Corpus contains read speech from 6 French, 6 English and 6 Zurich German speakers of Standard German. Speakers read 18 Standard German sentences (see Appendix A). To create perception experiments, 9 sentences per speaker were chosen such that sentence sets differed between speakers. Each of the 18 sentences appeared 6 times in the experiments: 3 times read by a French and 3 times read by an English speaker (see Chapter 4, Table 2). These 108 French- and English-accented sentences (2 accents × 6 speakers × 9 sentences) were signal-manipulated in different ways for the different perception experiments presented in Chapters 4, 5, and 6. The patterns of temporal variability in this material, along with the corresponding material from Zurich German speakers, are investigated from a speech production point of view in Chapter 3.

The Non-native Speaker-Individuality Corpus is a subset of the TEVOID corpus (Dellwo et al., 2012, 2015; Leemann et al., 2014) that was designed for the study of speaker-individual characteristics. The TEVOID speaker group is homogeneous in terms of native dialect, age, education, and second languages spoken, and speakers produced speech under various conditions. Our subset contains read speech from 16 Zurich German speakers in their native Zurich German as well as in their non-native French and English. Speakers read a list of 16 Zurich German, 16 French and 16 English sentences. French and English sentences were literal translations of the Zurich German sentences (see Appendix A). These 768 sentences (3 languages × 16 speakers × 16 sentences) were used for an experiment on speaker-idiosyncratic pausing behavior presented in Chapter 7 and for the evaluation of speakers' accent strength presented in Chapter 8. This corpus has now been enhanced in order to contain 48 sentences per speaker and language, though two of the 16 initial speakers involved in the experiment presented in Chapter 7 could not come back for additional recordings and were replaced by two further Zurich German speakers with

the same gender, age, educational level, and linguistic biography profile. An experiment where we apply some of the temporal measures presented in Chapter 3 to this enhanced corpus is currently in preparation (see Chapter 10.2).

# Part II

# Experimental investigations

# 3

# Temporal and rhythmic features in German as spoken by French, English and Zurich German speakers

This chapter presents an experiment that was carried out to describe patterns of temporal variability in the Non-native Speaker Origin Corpus. We applied a number of temporal measures to the speech material that was used for the perception experiments described in Chapters 4, 5, and 6:

- ▷ 2 measures of articulation rate and 2 measures of pausing;
- ▷ 14 rhythm metrics.

The main results of the present experiment are the following:

- ⇒ Native speech exhibited a higher articulation rate as well as fewer and shorter pauses than non-native speech.
- ⇒ Non-native speech revealed less durational variability in voiced and vocalic intervals, and more durational variability in unvoiced intervals than native speech.
- ⇒ French speakers of German tended to exhibit a higher proportion of the utterance over which speech is voiced than English speakers of German.
- ⇒ To a lesser extent, French and English speakers of German differed in the proportion of the utterance over which speech is vocalic as well as in the durational variability of voiced and consonantal intervals.

As this corpus was used to create stimuli for the perception experiments described in Chapters 4, 5, and 6, the patterns of temporal variability that were found to differ between foreign accents in the present experiment could be investigated as acoustic correlates of the perceptual results (see Chapter 6). Temporal features that were found to differ between native and non-native speech may reflect general properties of non-native speech and are discussed in Section 9.3.

# Introduction

To explore whether and how temporal patterns differed between the two non-native accents and the native accent of the Non-native Speaker Origin Corpus, we compared articulation rate, pausing behavior and a number of rhythm metrics between French-accented, English-accented and native German natural speech. Between-accent differences in acoustic measures of temporal variability point to possible cues that listeners may have used to complete the perceptual accent identification tasks in Chapters 4, 5, and 6. A brief overview of the literature on temporal measures of non-native speech in other corpora is given in Section 2.1.2 above. As the present experiment was designed to complement and possibly explain the perceptual results, we do not necessarily aim for a comprehensive discussion of its results *vis-à-vis* this literature.

# Materials and methods

## Speech material

For this experiment, we used the French-accented, English-accented, and native German speech material described in Section 2.2 as well as in Chapters 4, 5, and 6. It is important to note that a perceptual rating of the 108 non-native sentences by native listeners did not reveal differences between the French- and the English-accented sentences in terms of accent strength (see Chapter 4, Section 3.2 for details). This means that between-accent differences cannot be attributed to non-native accent strength.

## Annotation

To prepare the data for temporal measurement, the 108 non-native sentences and their native German counterparts were segmented by a trained phonetician (the author) using Praat (Boersma and Weenink, 2012). Segmentation and labeling decisions were based on visual inspection of waveforms and spectrograms as well as on auditory criteria. Silent pauses were annotated without the application of a particular duration threshold: pauses were labeled perceptually, with every silent part that was perceived as a pause labeled as such. Based on this segmentation, a tier was created containing consonantal and vocalic intervals (see Figure 3.1, tier 2). We further created a tier noting voiced and unvoiced intervals that were automatically calculated using the default pitch detection algorithm in Praat (tier 3). The last tier contained intervals between amplitude peaks, where the algorithm detected one peak per vocalic segment (tier 4; Dellwo et al., 2012).

## Temporal measures applied

We calculated two measures of articulation rate: *rateCV*, the number of consonantal and vocalic intervals per second (Dellwo, 2008) and *ratePeak*, the number of automatically detected peaks in the amplitude envelope per second. This roughly corresponds to the

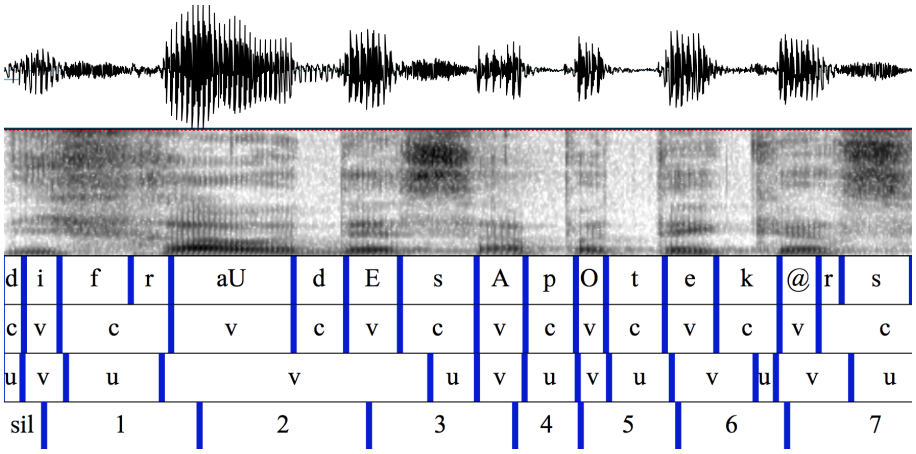



Figure 3.1: Spectrogram and waveform of a speech signal labeled for segments (tier 1), vocalic and consonantal intervals (tier 2), voiced and unvoiced intervals (tier 3) and amplitude-peak-to-amplitude-peak intervals (tier 4). The phrase reads *die Frau des Apothekers* 'the wife of the pharmacist'.

number of syllables per second (Mermelstein, 1975; Dellwo, 2008; Dellwo et al., 2012). We further applied two measures of pausing that are widely used in second language research: *pauseNbr*, the number of pauses, and *pauseDur*, pause durations (e.g. Trofimovich and Baker, 2006; de Jong et al., 2013; Künzel, 2013). Finally, we applied a number of rhythm metrics: metrics based on durational features of (a) vocalic and consonantal intervals; (b) voiced and unvoiced intervals; (c) intervals between automatically detected peaks in the amplitude envelope.

- Measures based on vocalic and consonantal interval durations:
  - *%V*, the percentage over which speech is vocalic (Ramus et al., 1999);
  - *varcoVln*, the rate-normalized standard deviation of vocalic interval durations (*varcoV*: White and Mattys, 2007a), calculated on log-transformed interval durations;
  - *nPVI_V*, the rate-normalized average difference between consecutive vocalic interval durations (Grabe and Low, 2002);
  - *varcoC*, the rate-normalized standard deviation of consonantal interval durations (Dellwo, 2006);
  - *nPVI_C*, the rate-normalized average difference between consecutive consonantal interval durations (Grabe and Low, 2002).
- Measures based on voiced and unvoiced interval durations:
  - *%VO*, the percentage over which speech is voiced (Dellwo et al., 2007);
  - *varcoVOln*, the rate-normalized standard deviation of voiced interval durations (*varcoVO*: Dellwo et al., 2007), calculated on log-stransformed interval durations;

  - *nPVI_VO*, the rate-normalized average difference between consecutive voiced interval durations (Dellwo et al., 2007);
  - *varcoUV*, the rate-normalized standard deviation of unvoiced interval durations (Dellwo et al., 2007);
  - *nPVI_UV*, the rate-normalized average difference between consecutive unvoiced interval durations (Dellwo et al., 2007).

- Measures based on interval durations between amplitude peaks:
  - *varcoPeak*, the rate-normalized standard deviation of interval durations between amplitude peaks (Dellwo et al., 2012);
  - *nPVI_Peak*, the rate-normalized average difference between consecutive interval durations between amplitude peaks (Dellwo et al., 2012).

As the distributions of vocalic and voiced intervals were strongly positively skewed, the measures *varcoV* (White and Mattys, 2007a) and *varcoVO* (Dellwo et al., 2007) were calculated based on log-transformed interval durations. Temporal measures were calculated sentence-by-sentence using the Praat plugin *Duration Analyzer* (written by Volker Dellwo; available at http://www.pholab.uzh.ch/en/leute/dellwo/software.html).

The calculation of *%V* and *%VO* is straightforward. For any given interval (*Int*), the calculation of *varcoInt* and *nPVI_Int* was calculated as follows:

$$varcoInt = 100 \cdot \frac{\Delta Int}{\overline{Int}},$$

where $\Delta Int$ denotes the standard deviation of interval durations and $\overline{Int}$ the mean interval duration.

$$nPVI\_Int = 100 \cdot \left[ \frac{\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right|}{m-1} \right],$$

where $d_k$ denotes the duration of the $k^{\text{th}}$ interval and $m$ the total number of intervals.

## Statistical analyses

Statistical analyses were performed using R software (R Core Team, 2014). To test for the effect of accent (with the levels *French-accented German*, *English-accented German* and *native German*) on each temporal measure, we constructed linear mixed effects models (LME) with *speakers' gender* and *accent* as fixed effects and *speaker* and *sentence* as random intercepts (R-package: *lme4*; Bates and Maechler, 2009). We tested effects by comparing a full model, which included the factor in question, to a reduced model, in which the factor was not included. Model comparison was performed using standard likelihood ratio tests (R-code: anova(full_model, reduced_model)). We report AIC (Akaike

Information Criterion) values for the relative goodness of fit of LMEs (Kliegl et al., 2011). For multiple comparisons we applied the Tukey method, using the R-package *multcomp*. We assumed an $\alpha$-level of 0.05.

# Results

Results on the effect of accent on durational patterns of speech are presented in Table 3.1. Considering the speaking rate measures, we obtained a significant effect of accent for *rateCV*, but not for *ratePeak*. Multiple comparisons revealed that *rateCV* was significantly different only between French-accented and native German. However, the three accents differed in speaking rate descriptively with native German showing the highest *rateCV* (M=9.91, SD=1.04), followed by English-accented German (M=8.89, SD=0.97) and French-accented German (M=8.19, SD=1.08; see Figure 3.2). A similar pattern was observed for *ratePeak*.

| Temporal measure | Factor | Result | | |
|---|---|---|---|---|
| *varcoUV* | *accent* | $\chi^2(2)$=26.24, | AIC=-103.22, | p<0.001* |
| *nPVI_UV* | *accent* | $\chi^2(2)$=22, | AIC=1627, | p<0.001* |
| *pauseNbr* | *accent* | $\chi^2(2)$=16.86, | AIC=557.12, | p<0.001* |
| *rateCV* | *accent* | $\chi^2(2)$=11.27, | AIC=428.49, | p<0.01* |
| *nPVI_V* | *accent* | $\chi^2(2)$=10.13, | AIC=1381, | p<0.01* |
| *pauseDur* | *accent* | $\chi^2(2)$=8.72, | AIC=49.76, | p<0.05* |
| *varcoVOln* | *accent* | $\chi^2(2)$=7.88, | AIC=-226.79, | p<0.05* |
| *nPVI_Peak* | *accent* | $\chi^2(2)$=5.37, | AIC=1440.80, | p=0.07 |
| *ratePeak* | *accent* | $\chi^2(2)$=5.16, | AIC=322.79, | p=0.08 |
| *%VO* | *accent* | $\chi^2(2)$=4.94, | AIC=1153.70, | p=0.08 |
| *varcoC* | *accent* | $\chi^2(2)$=4.46, | AIC=-381.44, | p=0.11 |
| *nPVI_VO* | *accent* | $\chi^2(2)$=3.93, | AIC=1543.8, | p=0.14 |
| *%V* | *accent* | $\chi^2(2)$=2.88, | AIC=969.75, | p=0.24 |
| *varcoPeak* | *accent* | $\chi^2(2)$=2.51, | AIC=-371.24, | p=0.29 |
| *varcoVln* | *accent* | $\chi^2(2)$=1.49, | AIC=-720.85, | p=0.47 |
| *nPVI_C* | *accent* | $\chi^2(2)$=0.09, | AIC=1401.20, | p=0.96 |

Table 3.1: Summary of statistics for temporal measures tested on French-accented, English-accented and native German speech. Acoustic measures are ordered according to magnitude of effect.

We found a significant effect for both the number (*pauseNbr*) and duration (*pauseDur*) of pauses; multiple comparisons revealed that native speakers produced significantly fewer and shorter pauses than non-native speakers. French speakers did not differ from English speakers, though (see Figure 3.2).

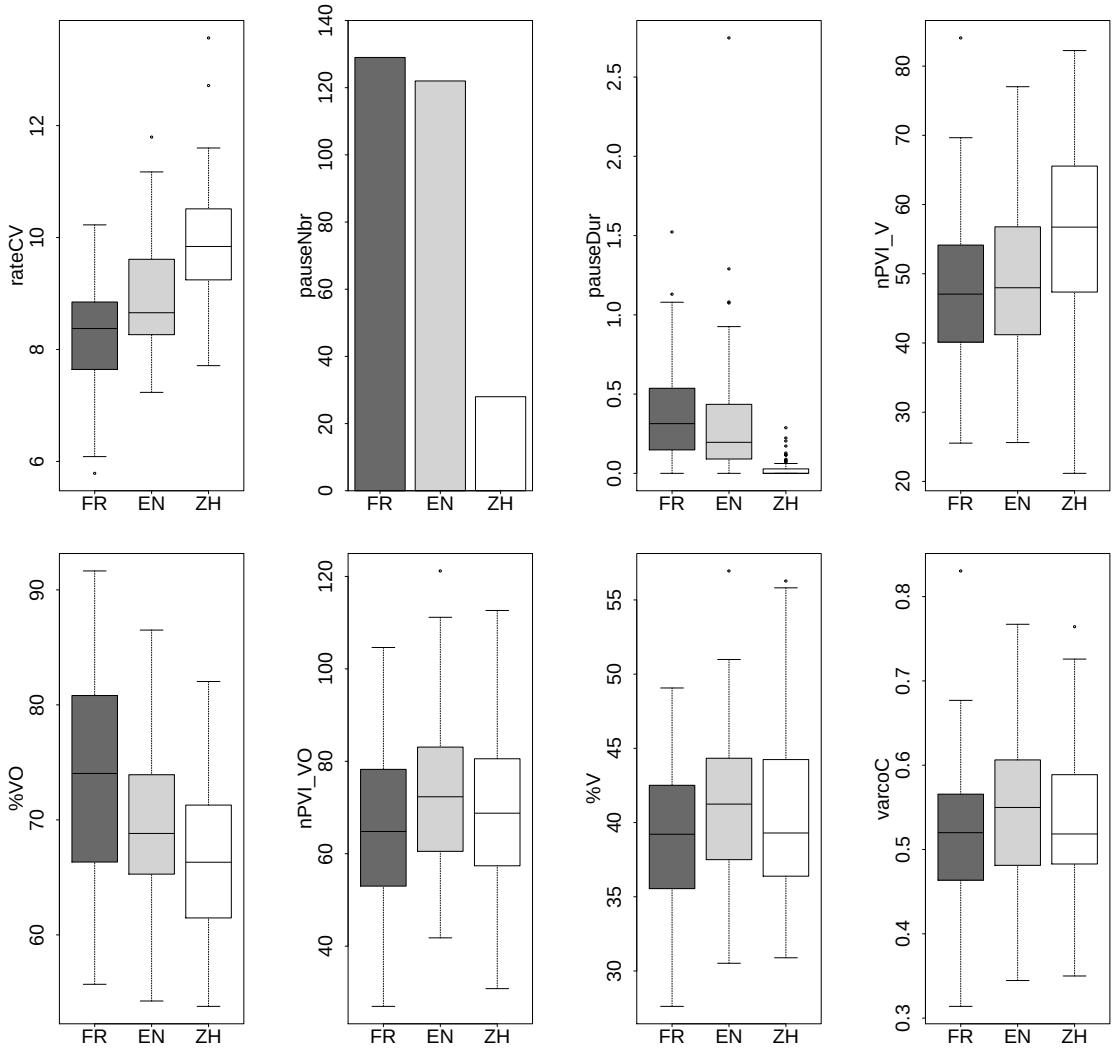



Figure 3.2: Boxplots of *rateCV*, *pauseNbr*, *pauseDur*, *nPVI_V*, *%VO*, *nPVI_VO*, *%V*, and *varcoC* for French-accented (left, dark gray), English-accented (center, light gray), and native German (right, white).

As for the rhythm metrics applied, we obtained a significant effect of accent for *nPVI_V*, *varcoVOln*, *varcoUV*, and *nPVI_UV*. Multiple comparisons revealed a significant difference between native and non-native accents for these measures, but not between the two non-native accents. Non-native speech revealed significantly lower interval duration variability than native speech in the vocalic and voiced measures, but significantly higher variability for the unvoiced measures.

We found no significant differences between French- and English-accented German. The highest descriptive differences between the two non-native accents were observed in the measures *%VO*, *nPVI_VO*, *%V*, and *varcoC*. Figure 3.2 illustrates that French speakers tended to produce a higher proportion of voiced intervals (*%VO*; M=73.95, SD=9.17) than

English (M=69.67, SD=6.91) speakers of German. On the other hand, French (M=66.09, SD=17.67) speakers of German showed lower values for *nPVI_VO*, than English speakers (M=72.78, SD=17.39). As for the percentage over which speech is vocalic, *%V*, English speakers (M=41.28, SD=5.43) exhibited higher values than French speakers (M=39.40, SD=4.90) of German. Finally, English speakers produced more variable consonantal interval durations (*varcoC*, M=0.55, SD=0.10) than French (M=0.52, SD=0.09) speakers.

# Discussion and conclusion

We found native speech to differ from non-native speech in terms of articulation rate and pausing behavior, where native speech was characterized by a higher articulation rate as well as fewer and shorter pauses. This replicates results reported by research on second language fluency (Trofimovich and Baker, 2006; Derwing et al., 2009; de Jong et al., 2013). Furthermore, we found non-native speakers to exhibit lower values for *nPVI_V*, therefore producing adjacent vocalic intervals with more similar durations than native speakers. This may reflect a tendency for non-native speakers of German to produce full vowels in contexts where native speakers reduce vowels, which is typically the case in unstressed syllables (Dauer, 1983). A similar result has been observed by Ordin and Polyanskaya (2015): French and German learners of English exhibit lower values compared with native speakers of English for the durational variability of vocalic intervals, but their variability increases as they advance in proficiency. Furthermore, our French and English speakers of German produced less durational variability in voiced intervals and more durational variability in unvoiced intervals than native speakers, which may be interpreted with regard to phenomena of elision or epenthesis to some extent (see Chapter 6 and Section 9.3). We found non-native speakers not to differ from native speakers in the percentage over which speech is vocalic, *%V*, which is unexpected considering findings by White and Mattys (2007b) and Dellwo (2010), reported in Chapter 2.

None of the 16 temporal measures applied revealed significant differences between French- and English-accented German. The two accents did, however, differ on a descriptive level in the percentage over which speech is voiced, *%VO*, with French speakers of German exhibiting higher values than English speakers. This may be due to interference from speakers' native languages: for example, French is characterized by shorter voice onset time than English, and this feature is typically influenced by speakers' native language in non-native speech (Hazan and Boulakia, 1993; Fowler et al., 2008). More generally, Dellwo et al. (2007) found that *%VO* is higher in French than in English speech, which also supports the hypothesis that interference from the native language plays a role here. To a lesser extent, French speakers also exhibited lower values than English speakers in the durational variability of adjacent voiced intervals, in the percentage over which speech is vocalic, and in the durational variability of consonantal intervals. We therefore conclude that features such as the ones captured by *nPVI_VO*, *varcoC*, *%V*, and, in particular, *%VO*, may be of use for perceptual decisions on non-native speakers' origin, in cases where these are not affected by signal manipulation (see Chapters 4, 5, and 6).

# 4

# Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition

This chapter contains a reprint of the paper: Kolly, M.-J., Dellwo, V. (2014). Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition. *Journal of Phonetics*, 42: 12–23.[1]

In this paper we present a series of perception experiments that examined listeners' ability to identify speaker origin in French- and English-accented German based on primarily temporal information. As French- and English-accented German were shown to differ in a number of temporal features in Chapter 3, we assume that time domain information contributes to listeners' accent identification performance. Stimuli for perception experiments were therefore heavily degraded in the frequency domain in order to contain different types of predominantly temporal features:

- ▷ Noise vocoded speech was manipulated to contain primarily amplitude envelope temporal features. 6-band noise vocoded speech contained more detailed frequency domain information than 3-band noise vocoded speech, and segment durations as well as information on voicing were absent from the stimuli.
- ▷ 1-bit requantized speech was manipulated to convey mainly segment durations and voicing temporal features; amplitude envelope information was absent from the stimuli.
- ▷ Monotonized *sasasa*-speech exposed listeners to voicing temporal patterns, as every voiced interval was turned into [a] and every unvoiced interval into [s]. Segment durations and amplitude envelope temporal information were absent from the stimuli.
- ▷ Natural speech was used as a control condition.

[1]DOI: http://dx.doi.org/10.1016/j.wocn.2013.11.004.

The main findings reported in this paper are the following:

⇒ Listeners identified foreign accents above chance based on primarily time domain information in 1-bit requantized and in 6-band noise vocoded speech. When no frequency domain information was available, e.g., in monotonized *sasasa*-speech, accent identification was no longer possible.

⇒ Listeners seemed particularly sensitive to segment durations: 1-bit requantized speech allowed for higher accent identification performance than 6-band noise vocoded speech.

⇒ As frequency domain information was reduced, accent identification performance decreased.

⇒ French-accented German was identified with higher performance than English-accented German in natural and 1-bit requantized speech.

Given these results, we put forward two possible conclusions: on the one hand, we suggested that the temporal structure of speech may be particularly relevant to speech perception in situations where frequency domain features are strongly degraded — as, for instance, in a noisy environment or on the telephone. In such situations, it may be that listeners can process the remaining frequency domain cues because they occur at specific and expected moments in the time domain. This hypothesis was later tested in Chapter 6. On the other hand, we discussed the (un-)naturalness of the stimuli as a possible limitation of this study. The signal types presented to listeners were not representative of everyday conversational situations or of acoustic impressions that listeners may encounter in natural environments. We suggested that time domain information may be sufficient for listeners to identify foreign accents if the signal type presented is one that occurs in natural environments. This hypothesis was tested in Chapters 5 and 6.

Journal of Phonetics 42 (2014) 12–23

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/phonetics

# Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition

Marie-José Kolly *, Volker Dellwo

*University of Zurich, Phonetics Laboratory, Department of General Linguistics, Plattenstrasse 54, 8032 Zurich, Switzerland*

ARTICLE INFO

*Article history:*
Received 24 January 2013
Received in revised form
26 October 2013
Accepted 9 November 2013

ABSTRACT

Foreign-accented speech typically contains information about speakers' linguistic origin, i.e., their native language. The present study explored the importance of different temporal and rhythmic prosodic characteristics for the recognition of French- and English-accented German. In perception experiments with Swiss German listeners, stimuli for accent recognition contained speech that was reduced artificially to convey temporal and rhythmic prosodic characteristics: (a) amplitude envelope durational information (by noise vocoding), (b) segment durations (by 1-bit requantisation) and (c) durations of voiced and voiceless intervals (by *sasasa*-delexicalisation). This preserved mainly time domain characteristics and different degrees of rudimentary information from the frequency domain. Results showed that listeners could recognise French- and English-accented German above chance even when their access to segmental and spectral cues was strongly reduced. Different types of temporal cues led to different recognition scores – segment durations were found to be the temporal cue most salient for accent recognition. Signal conditions that contained fewer segmental and spectral cues led to lower accent recognition scores.



## 1. Introduction

Foreign-accented speech contains numerous cues about the native language (L1) of its speakers (Cunningham-Andersson & Engstrand, 1987). If, for example, we consider Swiss-German- or French-accented English, it is typically easy for listeners who are familiar with these varieties to recognise these two accents. What are the acoustic cues for this? On a segmental level, for example, consonants may be pronounced at a different place of articulation, in a different manner of articulation, or with different degrees of voicing (see Leemann, 2011; Schmid, 2012a): the consonant in English *the* is likely to be pronounced [z] in a prototypical French accent, [d̥] in a prototypical Swiss German accent, thus differing from the English target [ð] in its place of articulation (French) or in place, manner and voicing (Swiss German). Similarly, /r/ in foreign-accented *random* is typically realised as a uvular trill [ʀ] or fricative [ʁ] by French speakers, as an alveolar trill [r] by Swiss German speakers. The first vowel in *random* would typically be nasalised ([ã]) by French and non-nasalised ([æ]) by Swiss Germans. Thus, segmental cues seem to play a large role for the recognition of these foreign accents (e.g. Cunningham-Andersson & Engstrand, 1987; Koster & Koet, 1993; Boula de Mareüil, Vieru-Dimulescu, Woehrling, & Adda-Decker, 2008; Park, 2013).

Apart from segmental cues there has also been a strong interest in prosodic phenomena of second language (L2) speech (Anderson-Hsieh, Johnson, & Koehler, 1992; Boula de Mareüil & Vieru-Dimulescu, 2006; Jilka, 2000; Magen, 1998; Munro, 1995; Munro, Derwing, & Burgess, 2010; Tajima, Port, & Dalby, 1997; Trouvain & Gut, 2007). This research typically deals with acoustic correlates of foreign accent degree, intelligibility, or foreign accent detection (temporal characteristics: Bent, Bradlow, & Smith, 2008; Dellwo, 2010; Holm, 2008; Munro & Derwing, 2001; Quené & van Delft, 2010; Tajima et al., 1997; Winters & O'Brien, 2013). However, the question whether particular foreign accents can be recognised based on specific prosodic cues has barely been tapped into. So far, it has been shown that speaker origin can be recognised in natural L2 speech (Derwing & Munro, 1997; Boula de Mareüil et al., 2008; Guntern, 2011; Kolly, 2013; Kumpf & King, 1997), in L2 speech with monotone intonation (Van Els & De Bot, 1987), in resynthesised L2 speech containing cues to intonation and segment durations only (Boula de Mareüil & Vieru-Dimulescu, 2006), but not in lowpass filtered L2 speech below 350 Hz (Van Els & De Bot, 1987). This body of research thus demonstrates that foreign accents can be recognised based on a variety of prosodic and segmental cues.

Little, however, is known about the role of time domain cues such as suprasegmental timing phenomena or speech rhythm in foreign accent recognition. Moreover, the lowpass filtering study by Van Els & De Bot (1987) suggests that after heavy reduction of frequency domain cues, foreign

* Corresponding author. Tel.: +41 44 634 59 48.
*E-mail addresses:* marie-jose.kolly@pholab.uzh.ch, mj.kolly@gmx.ch (M.-J. Kolly), volker.dellwo@uzh.ch (V. Dellwo).

0095-4470/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.wocn.2013.11.004

accent recognition is no longer possible. Somehow contradictory evidence can be found in the domain of L1 dialect recognition where lowpass filtered speech with a cutoff frequency of 250 Hz allows for the recognition of Swiss German dialects (Leemann & Siebenhaar, 2008). The same is true for lowpass filtered speech with an unknown cutoff in recognising English dialects (Bush, 1967). Furthermore, temporal cues like durations of consonantal and vocalic intervals allow listeners to discriminate between English dialects (White, Mattys, & Wiget, 2012). We take this as an indication that temporal cues may also play a role in the recognition of foreign-accented speech. The principal aim for the present study is to explore whether temporal characteristics of foreign accented speech are perceptually salient, by investigating how the reduction of listeners' access to segmental and spectral content of speech affects their ability to recognise foreign accents.

Why temporal characteristics? It is widely acknowledged that languages (Abercrombie, 1967; Grabe & Low, 2002; Pike, 1945; Ramus, Nespor, & Mehler, 1999) and dialects (Ferragne & Pellegrino, 2004; Leemann, Dellwo, Kolly, & Schmid, 2012; Schmid, 2012b; White et al., 2012; White & Mattys, 2007b) differ in their suprasegmental temporal organisation, or speech rhythm. Whether and to what degree language-specific rhythm allows for a classification of languages into rhythmic classes is a matter of heavy debate in the literature (see Arvaniti, 2012). However, there is strong evidence that languages can be discriminated based on auditory rhythmic characteristics (Ramus & Mehler, 1999; Ramus et al., 1999). Such characteristics have been associated with the sound of a Morse-code signal for some languages (e.g. English, German, Dutch) and with the sound of a machine-gun for others (e.g. French, Italian, Spanish; Lloyd James, 1929), while the latter expresses more regular rhythmic timing – in French as opposed to English, for example. In fact, there is evidence that durational characteristics of consonantal and vocalic intervals are perceived as more regularly timed in French than they are in English (Dellwo, 2008). It was also found that Mandarin speakers produce more regularly timed speech when speaking in synchrony while such effects cannot be obtained for English (Cummins, Li, & Wang, 2013). In summary, there is evidence for some languages to be more regularly timed than others, in speech production as well as in speech perception research.

Are rhythmic characteristics transferred from L1 to L2 speech? The literature demonstrates that this is true for some L1/L2-pairs and some durational variables, but not for others. For example, the rate-normalised durational variability of vocalic intervals[1] locates L2 speech in between the native and the target language values for rhythmically regular Spanish vs. irregular English (Carter, 2005; Gutiérrez Díez, Dellwo, Gavaldà, & Rosen, 2008; White & Mattys, 2007a). English and Dutch, which are both rhythmically irregular, show very similar values for native and target language as well as L2 speech (White & Mattys, 2007a). This points in favour of an L1-transfer hypothesis. However, other findings do not support such a hypothesis: Regarding the percentage over which speech is vocalic,[2] English learners of Spanish (White & Mattys, 2007a) as well as German learners of French and English (Dellwo, 2010) overshoot the values of their native as well as their target language. A high percentage over which speech is vocalic seems to be a general property of L2 speech. In fact, L2 speakers tend to lengthen the duration of vowels, particularly of unstressed vowels, giving the auditory impression of more regular speech timing (Adams & Munro, 1978; Taylor, 1981). Thus, L2 speech seems to be influenced by L1 durational characteristics for some variables and language pairs; other variables and language pairs, however, seem to reflect general properties of L2 speech rather than specific L1-transfer, as suggested by Taylor (1981) and Dellwo (2010). It therefore remains unclear, for L2 speech, which of the durational characteristics associated with speech rhythm are L1-specific, and which are a general feature of (L1-independent) L2 speech. It further remains widely unclear whether such acoustic variability between L2 accents is perceptually salient. While perceptually salient rhythmic differences between some languages have been empirically attested by different studies (Nazzi, Bertoncini, & Mehler, 1998; Ramus & Mehler, 1999; Ramus, Dupoux, & Mehler, 2003), the idea that such characteristics also play a role in L2 speech has been investigated empirically only for speech production (Dellwo, 2010; White & Mattys, 2007a, 2007b).

Durational characteristics of foreign-accented speech may be perceptually salient typically if speakers were to transfer durational patterns from a rhythmically more regular L1 to a less regular L2. This can be tested, for example, with French- and English-accented German speech: two foreign accents that stem from two languages that have been shown to differ in time domain characteristics (French and English; Abercrombie, 1967; Dellwo, 2006; Grabe & Low, 2002; Pike, 1945; Ramus et al., 1999). A rationale for this is the following: English and German, in contrast to French, are characterised by vowel reduction, complex syllables and consonant clusters, high durational variability between stressed and unstressed syllables. In comparison, French has less vowel reduction, less complex syllables and consonant clusters as well as less durational variability between stressed and unstressed syllables (Dauer, 1983; Auer, 2001). The percept of rhythmic regularity in French may be a result of such phonological characteristics. If language-typical phonological characteristics were indeed transferred from L1 to L2 speech, one would expect French accented German to sound rhythmically more regular than English accented German.

Cues for the perception of speech rhythmic characteristics are assumed to lie in the more or less regular recurrence of perceptually salient speech intervals. Since durational patterns are encoded on many levels in the speech signal, different types of such speech intervals have been considered to be acoustic correlates of speech rhythm: interstress intervals and syllables (Pike, 1945; Abercrombie, 1967), consonantal and vocalic intervals (Ramus et al., 1999), voiced and voiceless intervals (Dellwo, Fourcin, & Abberton, 2007; Fourcin & Dellwo, 2009), intervals related to amplitude envelope timing (Lee & Todd, 2004; Dellwo, Leemann, & Kolly, 2012; Tilsen & Johnson, 2008) or to fundamental frequency (Kohler, 2009). In research on speech perception, a small number of speech intervals have been used to study language discrimination based on durational characteristics: It has been shown that listeners can discriminate a rhythmically regular from an irregular language based on monotone lowpass filtered speech below 180 Hz (den Os, 1988) and based on the durational variability of consonantal and vocalic intervals in monotone *sasasa*-speech[3] (Ramus et al., 2003). Research on rhythm production and perception has thus mainly focused on temporal characteristics of vocalic and consonantal intervals. To test foreign accent recognition in conditions of heavily reduced frequency domain information, it thus seems reasonable to use different types of speech intervals to present time domain information to listeners. Durational characteristics of some speech intervals may contain more or less information about the L1 origin in L2 speech, which may lead to different accent recognition scores.

Based on the ideas presented above, we formulated the following research questions: To what degree can we reduce frequency domain characteristics of the speech signal such that listeners can still recognise two different foreign accents? And which type of temporal cue (i.e., which type of temporally structured speech interval) leads to higher accent recognition scores? To test this, Swiss German listeners were asked to recognise French- and English-accented German in signal-degraded speech containing primarily durational cues. In a between-subject design we used three different types of signal-degraded speech to provide listeners with different types of temporal cues. By doing this, we gain insight into the speech intervals that contribute more or less to accent recognition, i.e., the speech intervals which are (a) subject to durational L1-transfer

[1] A variable that has been shown to discriminate between hypothesised rhythm classes (Dellwo, 2006; White & Mattys, 2007a).

[2] This variable also discriminates between hypothesised rhythm classes (Ramus et al., 1999).

[3] Ramus & Mehler (1999) developed this procedure for delexicalisation, where the original utterances are resynthesised: every consonantal speech interval is replaced with the same [s]- and every vocalic speech interval with the same [a]-phone.

and (b) perceptually salient to the listeners regarding durations. To test listeners' attention to amplitude envelope durational cues (low frequency durational cues) we used noise vocoded speech; to test their attention to segmental durational information we used 1-bit requantised speech; to test their attention to the timing of the source signal (voice) we used *sasasa*-speech based on voiced and voiceless intervals (see Section 2). Furthermore, we degraded speech signals to different degrees to test whether listeners are sensitive to the reduction of spectral information (see Section 2). A between-subject design was used because listeners who are tested several times might improve between the conditions: it has been shown that distorted speech becomes more intelligible with experience (Licklider & Pollack, 1948, for 1-bit requantised speech; Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005, for noise vocoded speech). Since the signal degradations we applied render speech unintelligible, we presented the corresponding sentence transcript for each stimulus visually, which enabled listeners to parse the acoustic information to speech (Davis et al., 2005). In this way, they were able to process the temporal patterns and the potentially remaining spectral information in the signal. The following section explains the rationale for each signal degradation procedure and the type of temporal information as well as the amount of spectral information it contains.

## 2. Time domain cues in three types of signal-degraded speech

Signal degradation procedures were chosen in order to preserve different types of durational characteristics while severely reducing information in the frequency domain. Also, these signal degradation procedures reduced listeners' access to cues from the frequency domain to different degrees.

### 2.1. Noise vocoded speech

To obtain noise vocoded speech, amplitude envelopes are extracted from several frequency bands and used to modulate white noise in these frequency regions (Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995). The reduction of spectral information depends on the number of frequency bands used for amplitude envelope extraction: Fewer bands result in less spectral information (see Figs. 1 and 2). Cues to voicing are absent and cues to segment durations are severely degraded or absent. This type of signal thus displays time domain information in the form of amplitude envelope timing cues; however, the reduction of the number of frequency bands not only reduces spectral detail, it also reduces fine-grained temporal information, since amplitude timing cues are lost for these “missing” frequency bands. The perceptual impression of noise vocoded speech with a small number of frequency bands can be described as a succession of syllable beats in the form of white noise pulses to which the phenomenon of speech rhythm is most likely closely related (Cummins & Port, 1998; Lee & Todd, 2004; Tilsen & Arvaniti, 2013).

To judge the amount of spectral information that can be obtained from noise vocoded speech, we ran informal experiments, which showed that listeners cannot identify single phones in 6-band noise vocoded speech; however, they can discriminate noise vocoded vowels as well as different sibilants to some degree, given two categories as options. Thus rudimentary spectral information remains in 6-band noise vocoded speech (see Fig. 1). For 3-band noise vocoded speech, single phones could not be recognised nor could they be discriminated when category information was available. 3-band noise vocoded speech can thus be said to contain almost no spectral cues that might lead to the identification of individual segments (see Fig. 2 vs. Fig. 1). Similarly, noise vocoded sentences with a small number of frequency bands are unintelligible. However, listeners' access to a corresponding sentence transcript alleviates the processing of the temporal patterns as well as of the rudimentary frequency domain information contained in the signals (see Davis et al., 2005). Therefore, we presented listeners with sentence transcripts of the stimuli for two signal conditions: 6-band noise vocoded speech and 3-band noise vocoded speech. In a third signal condition we reduced listeners' access to spectral cues by presenting them 6-band noise vocoded speech without sentence transcripts.

### 2.2. 1-bit requantised speech

1-bit requantised speech was created by reducing the quantisation rate of the digital speech signal, originally 16-bit, to 1-bit by setting the amplitude value of every sample to one of two arbitrarily chosen quantisation levels: -1 (for sample amplitudes $<0$) and 0 (for sample amplitudes $>0$). 0 was included in the arbitrary choice to allow silences in the original signal to remain silent in the delexicalised signal. To exclude f0 influences, pitch

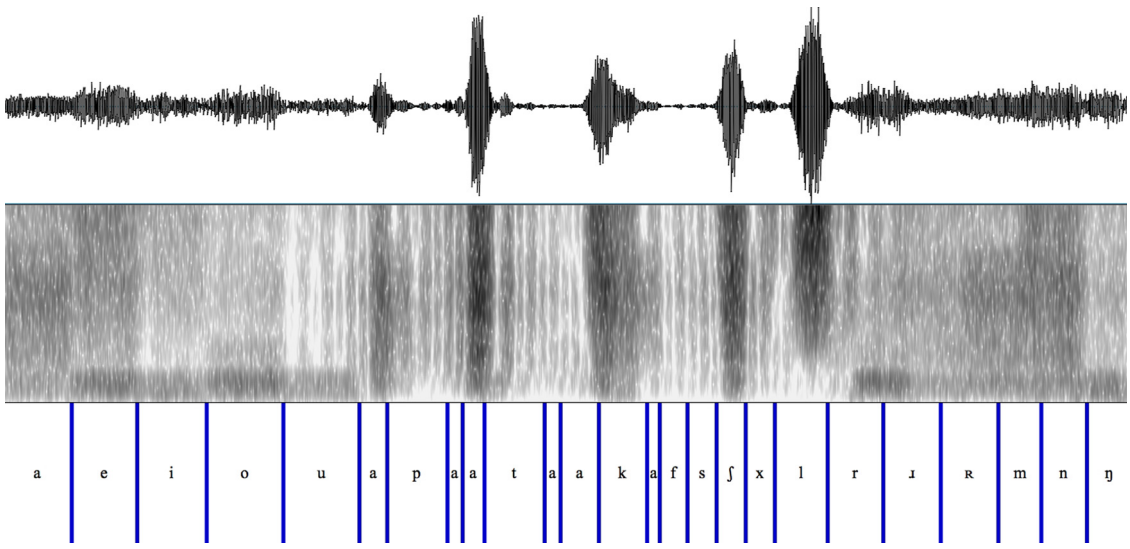



**Fig. 1.** Waveform and spectrogram for a number of different phones; 6-band noise vocoded speech.

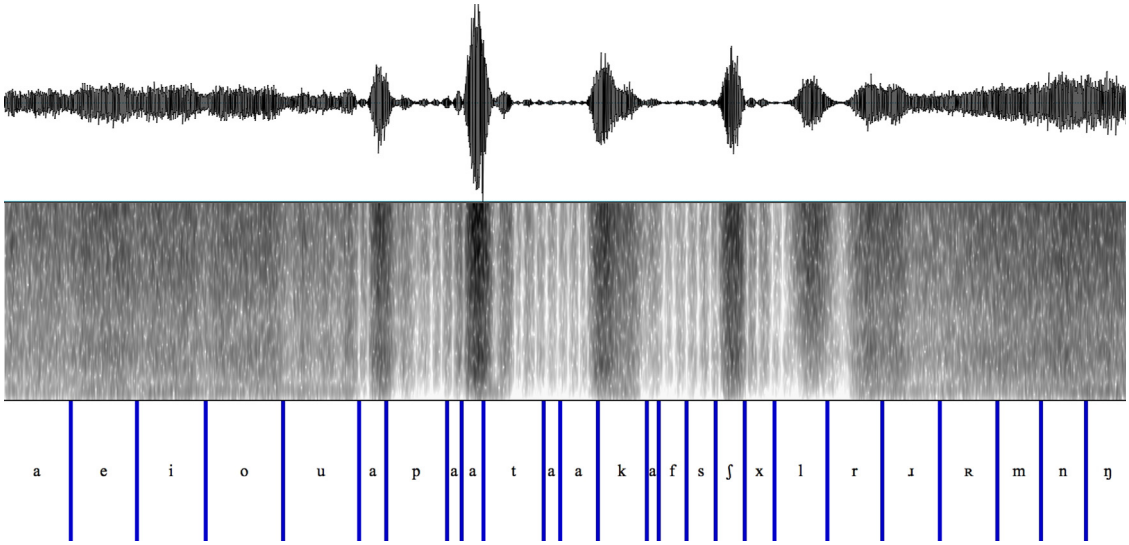



**Fig. 2.** Waveform and spectrogram for a number of different phones; 3-band noise vocoded speech.

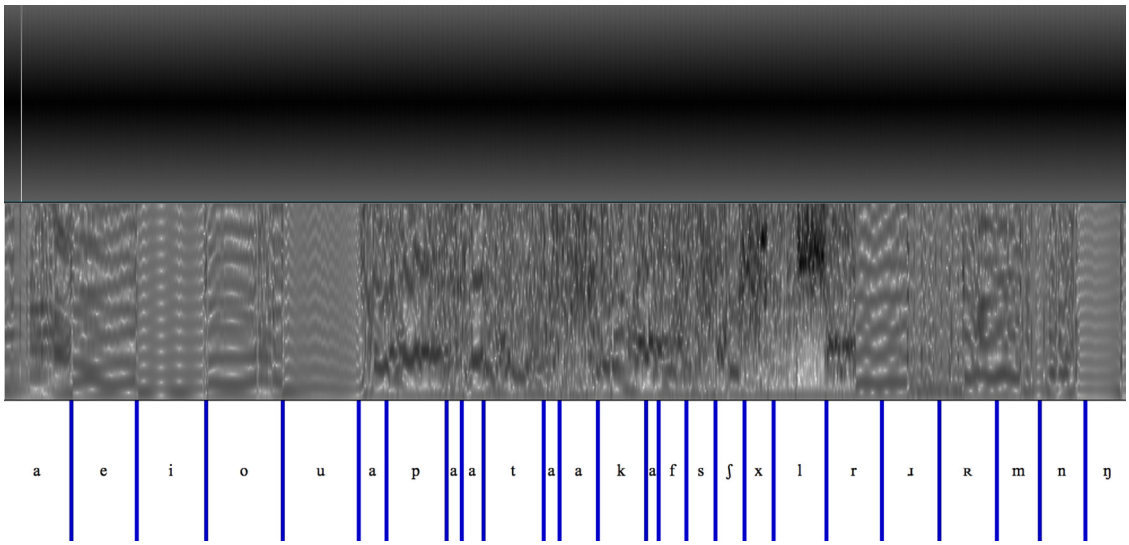



**Fig. 3.** Waveform and spectrogram for a number of different phones; 1-bit requantised speech.

was monotonised. This digital operation is similar and leads to nearly equal results as analogue methods of infinite peak clipping (Licklider & Pollack, 1948). 1-bit requantisation thus also severely degrades speech in the frequency domain while the boundaries of individual segments can typically be well obtained. This type of signal thus contains temporal information in the form of segment durations and voicing cues while the amplitude envelope is entirely lost (see Fig. 3).

Informal experiments with 1-bit requantised speech have shown that the two vowels [a] and [e] could be discriminated from any other vowel (given the categories as options). This was not the case for other vowel pairs. Also, the two sibilants [s] and [ʃ] could be discriminated, but no other pair of fricatives could. However, 1-bit requantised speech contains more manner cues: given the categories as options, plosives, fricatives and nasals could be told apart. 1-bit requantised speech thus contains more cues to the manner of articulation and voicing of consonants than 6-band noise vocoded speech (see Fig. 3 vs. Fig. 1).

### 2.3. *Sasasa-speech*

*Sasasa*-speech, where speech is delexicalised by turning all consonantal intervals into the same [s] and all vocalic intervals into the same [a], was developed by Ramus & Mehler (1999). In the present research we used this method but based it on voiced and voiceless speech intervals. We thus replaced every voiced interval (i.e., non-interrupted periodicity) with a pre-recorded [a] and every voiceless interval with a pre-recorded [s]. There are two reasons why we built *sasasa*-speech based on voiced and voiceless intervals instead of the more commonly used vocalic and consonantal intervals: First, the somewhat categorical and potentially problematic definition of e.g. approximants being consonantal is levelled out (see Wiget, White, Schuppler, Grenon, Rauch, & Mattys, 2010). Second, while durational measurements based on voiced and voiceless intervals discriminate languages similarly to measurements based on vocalic and consonantal intervals (Dellwo et al., 2007; Fourcin & Dellwo, 2009), the former distinction is perceptually salient even without any prior phonological knowledge about a language.

**Table 1**
Durational cues preserved by the signal degradation procedures

| | 1-bit requantised speech | Noise vocoded speech | *sasasa*-speech |
|---|---|---|---|
| **Amplitude envelope** | Absent | Present | Absent |
| **Segment durations** | Partly present | Absent | Absent |
| **Voicing cues** | Present | Absent | Present |
| ***Summary*** | *Segmental temporal cues* | *Amplitude envelope temporal cues* | *Voicing temporal cues* |

**Table 2**
Combination of 9 sentences per speaker.

| Speaker (French/English) | Sentences | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **01** | 01 | 07 | 13 | 02 | 08 | 14 | 03 | 09 | 15 |
| **02** | 02 | 08 | 14 | 03 | 09 | 15 | 04 | 10 | 16 |
| **03** | 03 | 09 | 15 | 04 | 10 | 16 | 05 | 11 | 17 |
| **04** | 04 | 10 | 16 | 05 | 11 | 17 | 06 | 12 | 18 |
| **05** | 05 | 11 | 17 | 06 | 12 | 18 | 07 | 13 | 01 |
| **06** | 06 | 12 | 18 | 07 | 13 | 01 | 08 | 14 | 02 |

Since *sasasa*-speech contains resynthesised segments, original segmental, spectral, f0 and amplitude information is completely absent from the signal (therefore, no figure of a *sasasa*-signal is displayed here). It is thus impossible to discriminate any pair of phones, unless they differ in terms of voicing contrast (e.g. [s] vs. [z]).

### 2.4. Summary

In conclusion, Table 1 shows a summary of which acoustic cues are preserved by the signal degradation procedures used in our experiments.

## 3. Materials and methods

### 3.1. Subjects

Our between-subject design involved six groups of 10 listeners per signal condition for a total of 60 subjects, all of which were native speakers of Swiss German dialects. Each condition was balanced for a similar number of male and female participants. The age of the subjects ranged between 19 and 34 years (mean = 25.08). None of the listeners reported any significant problems with hearing or sight. Most subjects were students from Zurich University, some were (former) students from other Swiss Universities. Thus, due to listeners' origin, age and educational level, they were assumed to have had a similar amount of contact with French and English native speakers respectively, as well as with foreign-accented German in general. However, subjects may have had more experience with French native speakers because there is a French speaking part in Switzerland. Subjects' basic school education in French and English was comparable.

### 3.2. Material

We collected speech from twelve speakers, six French and six English native speakers (three males and three females each). The French speakers were socialised and lived in the French speaking part of Switzerland (five in the canton of Fribourg, one in the canton of Vaud). The English speakers were socialised in the US or in Canada, one of the female speakers in the UK, and they lived in Switzerland at the time when they were recorded. The ages of the speakers ranged between 23 and 56 years (mean = 31.92). Their self-assessed proficiency in German ranged from B1 to B2 (intermediate) for the French speakers and from A1 to B2 (beginner to intermediate) for the English speakers (see Council of Europe, 2013). Speakers were rated for degree of foreign accent on a 5-point scale by Swiss German listeners, in an experimental condition involving natural speech (1 = very strong accent; 2 = strong accent; 3 = medium accent; 4 = slight accent; 5 = no accent). A two-sample *t*-test showed that there was no significant difference in accent degree between both non-native speaker groups ($t = -0.58$; ns; df = 5.86): the mean accent degree was 2.89 for French speakers and 2.70 for English speakers.

Speakers read a list of 18 German sentences (see Appendix A). Sentences were taken from a set of Italian materials used by Nazzi et al. (1998) and translated to German. Sentence length varied between twelve and 16 syllables. Prior to the recording, speakers familiarised themselves with the material by reading the sentences aloud. They were recorded in a quiet room at Zurich University or in their respective homes with a *Fostex FR-2LE* solid-state recorder (digitised with a sampling rate of 48 kHz and a quantisation rate of 16 bit) and a *Sennheiser MKE 2p-c* clip-on microphone. If filled pauses occurred during a sentence, speakers repeated the sentence spontaneously or, if not, they were asked to do so. Sentences with silent pauses were not repeated. A two-sample *t*-test showed that there was no significant difference in the number of pauses per sentence between both non-native speaker groups ($t = -1.62$; ns; df = 96.19): the mean number of pauses per sentence was 0.87 for French speakers and 1.22 for English speakers. There was, however, a significant difference in pause durations between the non-native speaker groups ($t = 3.03$; $p < 0.01$; df = 57.52): the mean pause duration was 0.29 s for French speakers and 0.20 s for English speakers. Nine sentences per speaker were chosen for the experiment to contain 108 stimuli. To avoid equal sentence sets for different speakers, we distributed three times 18 sentences among French and English speakers respectively in the way shown in Table 2 (see Appendix A for sentence numbers). Each of the 18 sentences appeared six times in the experiment: three times spoken by French and three times by English native speakers.

Manipulated stimuli were constructed using Praat signal processing software (Boersma & Weenink, 2012).[4] All stimuli were scaled to an intensity of 70 dB.

- To obtain noise vocoded stimuli, every sentence was first bandpass filtered between 50 Hz and 8000 Hz. This band was then divided into a certain number of logarithmically spaced frequency bands (six and three) by bandpass filtering. The cutoff frequencies for the six frequency bands were 50 Hz, 116.50 Hz, 271.44 Hz, 632.46 Hz, 1473.61 Hz, 3433.50 Hz and 8000 Hz. The cutoff frequencies for the three frequency bands were 50 Hz, 271.44 Hz, 1473.61 Hz and 8000 Hz. The same cutoff frequencies were used to filter white noise in order to obtain six and three noise bands respectively. The amplitude envelope was extracted from each speech band by half-wave rectification and lowpass filtering at 10 Hz. These amplitude envelopes were then multiplied with the corresponding noise bands and, finally, the modulated noise bands were summed to obtain a noise vocoded sentence (see Fig. 4).
- To obtain (monotone) 1-bit requantised stimuli, pitch was first monotonised: We replaced the pitch points of every sentence with the mean pitch value of the sentence. Subsequently, the amplitude of every sample of the signal was set to −1 (for sample amplitudes $<0$) or to 0 (for sample amplitudes $>0$). Therefore, the quantisation rate of the speech signal was converted to 1-bit (see Fig. 5).
- (Monotone) *sasasa*-stimuli were created with the Praat plug-in tool *Sasasa delexicaliser* (see footnote 4) and based on voiced and voiceless intervals. This tool requires annotated sound files (Praat TextGrids) with a tier coding voiced and voiceless parts of the signals. The latter were generated automatically using the pitch detection algorithm implemented in Praat. The *Sasasa delexicaliser* replaces voiceless intervals with a pre-recorded [s], voiced intervals with a pre-recorded [a] (male voice), preserving only interval durations from the original sounds (see Fig. 6).

### 3.3. Procedure

Listeners were tested individually, in a quiet room at Zurich University or in their own homes. They were presented with the 108 stimuli over high-quality earphones. The stimulus order was randomised separately for each subject. The experiment lasted between 15 and 25 min. Before the start of

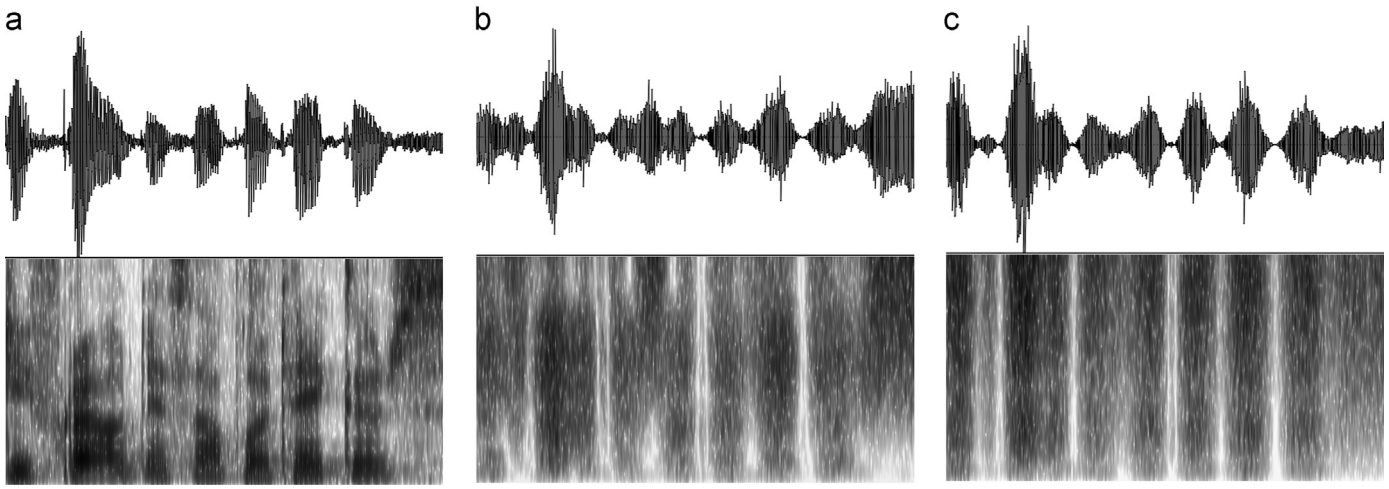



**Fig. 4.** *Die Frau des Apothekers* 'The wife of the pharmacist' spoken by a native English speaker; natural speech (a, left), 6-band noise vocoded (b, centre), 3-band noise vocoded (c, right).

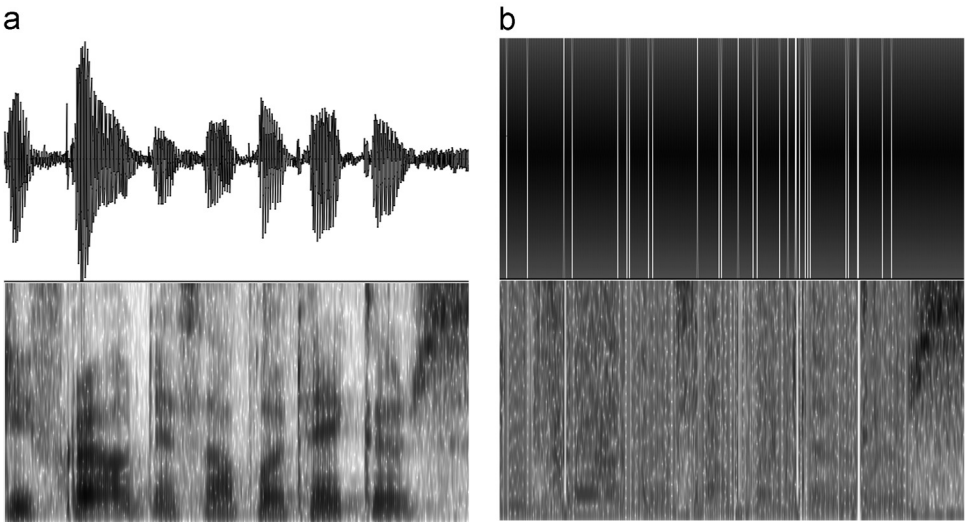



**Fig. 5.** *Die Frau des Apothekers* 'The wife of the pharmacist' spoken by a native English speaker; natural speech (a, left) and 1-bit requantised (b, right).

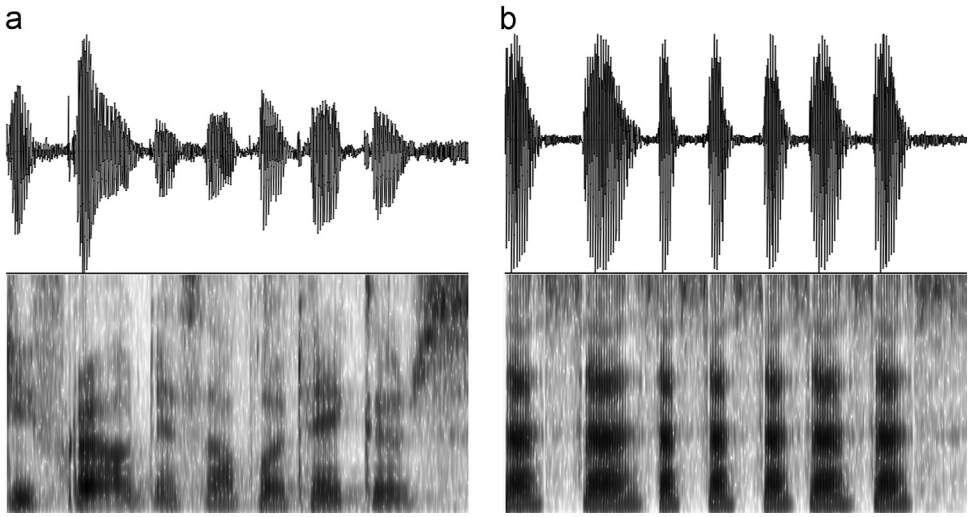



**Fig. 6.** *Die Frau des Apothekers* 'The wife of the pharmacist' spoken by a native English speaker; natural speech (a, left) and *sasasa*-delexicalised (b, right).

the experiment, listeners were familiarised with the experiment interface and with delexicalised speech through the presentation of two random stimuli from the signal condition they would be tested with. Sentence transcripts were given for four of the six conditions: 6-band noise vocoded speech, 3-band noise vocoded speech, 1-bit requantised speech and *sasasa*-speech. To this end, the sentence corresponding to the acoustic stimulus was presented on a laptop computer screen two seconds preceding the acoustic stimulus, and remained on the screen during the acoustic stimulus presentation. In these cases, listeners had access to lexical and syntactic information prior to and while listening to the delexicalised sentence. Two further conditions were tested without the presentation of sentence transcripts: another condition with 6-band noise vocoded speech, and natural speech. For each stimulus, listeners were asked to decide whether they heard French- or English-accented German. They were encouraged to respond intuitively. Also, they had to indicate the certainty of their response on a 3-point scale. Listeners responded by clicking on the corresponding button using an experiment interface (the Praat plug-in tool *Sentence presenter*) on a laptop computer.

### 3.4. Data analysis and statistics

Based on each listener's score we computed the response bias free measure $d'$ derived from signal detection theory (Green & Swets, 1966). Perfect sensitivity (i.e., perfect discrimination of two types of signals) starts at a $d'$-value of 4, and a $d'$-value of 0 indicates sensitivity at chance level.[5] Statistical analyses were performed using *R* (R Core Team, 2013). Since we did multiple comparisons on related data, we have chosen a conservative significance level with $\alpha = 0.01$.

## 4. Results

One-sample *t*-tests with $d'$ as a dependent variable show that accent recognition was significantly better than chance in several experimental conditions: In natural speech ($t = 15.04$; $p < 0.001$; df = 9), in 1-bit requantised speech ($t = 13.64$; $p < 0.001$; df = 9), and in 6-band noise vocoded speech ($t = 4.62$; $p < 0.001$; df = 10). The remaining conditions did not allow for a discrimination of the two signal types: 6-band noise vocoded speech without sentence transcripts ($t = 0.69$; ns; df = 12), 3-band noise vocoded speech ($t = 1.77$; ns; df = 9) and *sasasa*-speech ($t = -1.34$; ns; df = 9). These results are presented in Fig. 7.[6]

Fig. 7 further shows a decline in listener's ability to recognise accents, as frequency domain cues decrease in the different conditions (see Section 2). We computed a univariate ANOVA that shows a significant effect between conditions ($F(5, 58) = 91.61$; $p < 0.001$). Post-hoc tests reveal that all group comparisons are highly significant, except for comparisons between the conditions that did not enable accent recognition. Also, the 6-band noise vocoded condition does not significantly differ from the 3-band noise vocoded condition.

Furthermore, we computed the percentage of correct responses for the French- and English-accented stimuli separately (here it was not possible to compute $d'$, since we were interested in the responses to each of the two signal types). Two-sample *t*-tests with the percentage of correct responses as a dependent variable show that accent recognition scores are (marginally) significantly higher for French-accented stimuli in natural speech ($t = -2.83$; $p = 0.01$; df = 16.32) and significantly higher in 1-bit requantised speech ($t = -3.16$; $p < 0.01$; df = 16.09). Scores for both signal types do not significantly differ in 6-band noise vocoded speech ($t = -0.57$; ns; df = 20), 6-band noise vocoded speech without sentence transcripts ($t = 0.58$; ns; df = 22.70), 3-band noise vocoded speech ($t = 1.81$; ns; df = 15.75) and *sasasa*-speech ($t = -1.09$; ns; df = 17.94). These results are presented in Fig. 8.

[4] Praat scripts for delexicalisation and plug-in tools were written by the second author and are available at http://www.pholab.uzh.ch/leute/dellwo/software.html.

[5] To obtain $d'$, one has to compute the number of hits, false alarms, misses and correct rejections per subject. The value $d'$ is then given by the *Z*-value of the hit-rate minus the *Z*-value of the false-alarm-rate. An alternative to $d'$ is the non-parametric measure $A'$ (Donaldson, 1992). We also calculated $A'$ in addition to $d'$, but since we obtained proportionally equal results, we have not reported these values.

[6] A subset of this data has previously been published in a working paper (Kolly & Dellwo, 2013).

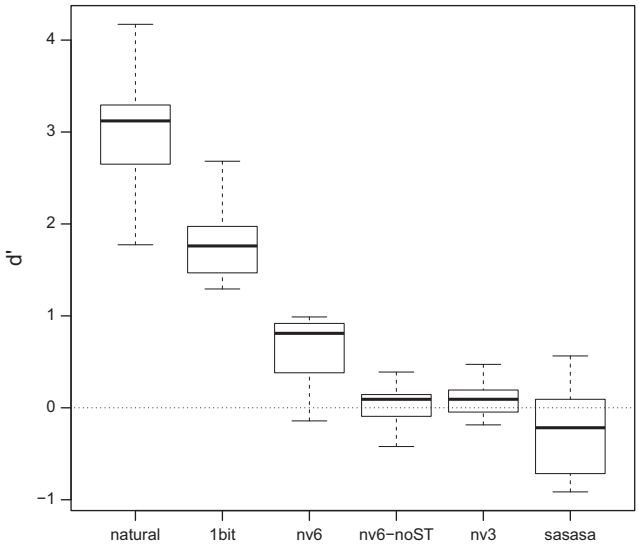



**Fig. 7.** Perceptual recognition of a French or English accent in German natural and delexicalised L2 speech. The dotted line indicates performance at chance (natural = natural speech; 1bit = 1-bit requantised speech; nv6 = 6-band noise vocoded speech; nv6-noST = 6-band noise vocoded speech without sentence transcripts, nv3 = 3-band noise vocoded speech; sasasa = *sasasa*-speech).

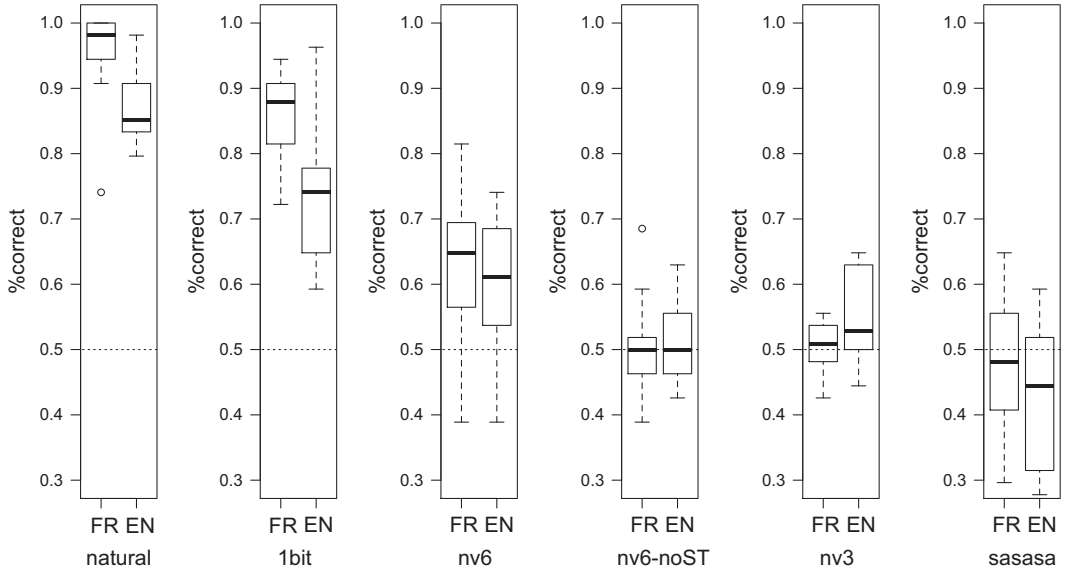



**Fig. 8.** Perceptual recognition of a French (left) or English (right) accent in German natural and delexicalised L2 speech. The dotted line indicates performance at chance (natural = natural speech; 1bit = 1-bit requantised speech; nv6 = 6-band noise vocoded speech; nv6-noST = 6-band noise vocoded speech without sentence transcripts, nv3 = 3-band noise vocoded speech; sasasa = *sasasa*-speech).

## 5. Discussion

The present experiments investigated the contribution of speech temporal cues to the recognition of foreign accents. We used different signal degradation procedures to present different types of time domain cues to our listeners. Furthermore, we used signal conditions that contain different degrees of frequency domain information. Two-alternative forced choice perception experiments with Swiss German listeners showed that French-accented German speech and English-accented German speech can be recognised above chance based on signal-degraded speech containing primarily temporal cues. This is the case for 6-band noise vocoded as well as for 1-bit requantised speech. Stimuli that contain less or no spectral information, such as 3-band noise vocoded speech and *sasasa*-speech, did not allow for accent recognition; neither did stimuli that allow less or no access to spectral information, namely 6-band noise vocoded speech without the presentation of sentence transcripts. It thus appears that the conditions where listeners have no access to spectral cues hinder their ability to recognise French- and English-accented German.

Depending on the perspective, there are two possible explanations for these results:

The findings reported in the present study can be explained by the fact that the different types of stimuli contain different types of time domain information. The primarily temporal cues in 1-bit requantised and in 6-band noise vocoded speech were sufficient to recognise French and English accents in German L2 speech: the absence of amplitude temporal cues in 1-bit requantised speech or the absence of periodicity and cues to segment durations in noise vocoded speech did not hinder accent recognition. However, the absence of amplitude temporal cues as well as cues to segment

durations in *sasasa*-speech did not suffice to recognise accents. Therefore, cues to segment durations in 1-bit requantised speech and cues to amplitude envelope timing in 6-band noise vocoded speech seem to be (a) subject to durational L1-transfer and (b) perceptually salient to listeners, since they enable accent recognition. Since 1-bit requantised speech was significantly better recognised than 6-band noise vocoded speech, listeners possibly rely more on segment durations (or on a combination of segment durations and voice timing, see Table 1) than on lower frequency cues to identify speaker origin when listening to French- and English-accented German speech. This result is in line with findings by Tajima et al. (1997), Holm (2008), and Quené and van Delft (2010) who observe that L2 speech becomes more intelligible to listeners if segment durations are manipulated to match the target language. We conclude that durational characteristics of segments as a relevant speech interval, possibly in combination with voicing, are highly salient characteristics of the foreign accents investigated in this study. The same holds, to a lesser degree, for amplitude envelope durational characteristics. Voiced and voiceless intervals alone (in *sasasa*-speech), however, do not seem to be perceptually salient speech intervals of the foreign accents at hand.

Alternatively, the findings reported above can be explained by the degree of frequency domain information conveyed in the different stimulus conditions. As presented in Section 2, the severely degraded spectral information in 1-bit requantised speech still contains more frequency domain information, in particular about the manner of articulation of consonants, than 6-band noise vocoded speech. This explains the higher accent recognition scores for listeners presented with 1-bit requantised speech in a different manner. In a more global view, Fig. 7 and the statistical results illustrate that the reduction of listeners' access to spectral cues results in lower accent recognition scores. The lesser degree of spectral information contained in 1-bit requantised speech as opposed to natural speech, in 6-band noise vocoded speech as opposed to 1-bit requantised speech, in the noise vocoded conditions with no access to spectral cues and in *sasasa*-speech as opposed to 6-band noise vocoded speech with sentence transcripts all led to a reduction of listeners' ability to recognise the foreign accents at hand. We conclude that spectral information is a very salient characteristic of foreign accents. This result is in line with the finding that lowpass filtering has a great detrimental effect on foreign accent perception: listeners' ratings of foreign accent degree based on natural and lowpass filtered stimuli do not correlate (Munro, 1995), and foreign accents could not be recognised in lowpass filtered speech below 350 Hz (Van Els & De Bot, 1987). Reducing frequency domain information thus affects listeners' ability to recognise foreign accents. The segmental information contained in 1-bit requantised speech or in 6-band noise vocoded speech is extremely rudimentary but still supports accent recognition above chance. The question thus arises: Is the interplay between time domain and frequency domain information necessary to solve the perceptual task at hand?

### 5.1. Interplay between time domain and frequency domain characteristics

Listeners may rely on the interplay between temporal and rudimentary spectral cues for accent recognition in our experiments. In fact, our results support a hypothesis presented in Dellwo (2010): It might well be that the temporal structure of the speech signal is crucial for listeners' ability to process this little amount of frequency domain information. Thus, the durational structure of speech might be essential to speech perception in those situations where spectral information is strongly degraded (e.g. in a noisy environment) – the remaining spectral cues can be processed by listeners because they occur at the expected moments in the time domain.

Accent recognition based on temporal characteristics alone was not possible with the stimuli and experimental design presented here. It seems evident, however, that temporal characteristics play a role in L2 speech, as shown by the accent recognition scores based on severely degraded speech in the present experiments. A small number of other studies report the perceptual importance of durational characteristics in L2 speech: Tajima et al. (1997), Quené and van Delft (2010), and Holm (2008) illustrate the relevance of segment durations for L2 speech intelligibility; Munro et al. (2010) show that the complete temporal distortion in randomly spliced backwards speech hinders listeners' ability to detect foreign accents.

A different experimental design might yield different accent recognition scores for stimuli containing temporal cues alone. For example, an extensive training of listeners with a second set of similar stimuli might raise the accent recognition scores. Also, an accent discrimination task, where listeners have to assign every stimulus to one of two reference stimuli (e.g. ABX task, see White et al., 2012) is likely to be easier to solve than the recognition task at hand: In such a design, listeners would not have to resort to their own experience of what e.g. French- or English-accented German sounds like – and could sound like in a severely degraded form. A possible limitation of this study lies in the naturalness of the stimuli: signal types such as 1-bit requantised, noise vocoded or *sasasa*-speech are not very representative of everyday communicative situations. One could think of alternative signal types that occur more frequently in natural environments and still maintain predominantly temporal characteristics, as for example lowpass filtered stimuli. The bandpass characteristics are similar to speech that is perceived under adverse conditions like someone talking in a different room with the door closed. In follow-up studies, it would be interesting to test the hypothesis that emerged in the present discussion: Time domain characteristics play an important role for speech perception when only little frequency domain information is available to the listener. This would shed more light on the interplay between cues from the time domain and the frequency domain in foreign accent perception.

### 5.2. French L1 vs. English L1

Additional results from the present experiments have shown that French-accented stimuli were recognised better than English-accented stimuli in natural as well as in 1-bit requantised speech. However, there were no such differences in the noise vocoded conditions and in *sasasa*-speech. Moreover, after taking part in the experiment, listeners often reported having had more ease in recognising the French-accented stimuli. Some listeners spontaneously said that the English-accented stimuli sounded more similar to native German. Given that the two non-native speaker groups did not differ in accent degree (see Section 3.2), a possible explanation for the better recognition of French accents is the fact that Swiss German subjects have arguably more experience with French speakers than with English speakers, since there is a French speaking part in Switzerland. French-accented German might thus have been more easily recognisable to these listeners. However, if this were true one would expect the effect to occur for all stimulus types where accents were recognised above chance – thus also for 6-band noise vocoded speech. Another tempting explanation is that the results and listeners' statements are in line with the speech rhythm transfer hypothesis proposed in Section 1: If French were, in fact, perceived as rhythmically more regular, and English and German as more irregular, French-accented German would be likely to sound rhythmically less native-like than English-accented German. Again, since the recognition scores do not differ for both accents in 6-band noise vocoded speech, an additional or alternative explanation must be sought in the frequency domain. The additional segmental and voicing information in natural and in 1-bit requantised speech, as compared to 6-band noise vocoded speech, may carry characteristics typical of a French accent – one may think of cues to manner and voicing, where French speakers would typically voice consonants that are usually voiceless in native German or in English (Neuhauser, 2011).

On the one hand, results about perceptually salient cues of foreign accented speech may have implications for the field of L2 acquisition: some speakers may wish to reduce their foreign accent in order to sound more native-like. For example, this might be the case if they are discriminated against because of their particular accent and origin (cf. Lippi-Green, 1997: 229). To sound more native-like, one needs to know

which acoustic characteristics of the foreign accent are perceptually salient to native listeners. The results of this study point to the fact that segment durations (possibly in combination with voicing) are salient temporal characteristics of French and English accented German. If French and English learners focus on German segment durations, their production of German vowel quantity, for example, might improve. However, more research is needed before such results can be applied in the L2 classroom. On the other hand, our research possibly has implications for the field of forensic phonetics, where experts assess (often incriminating) speech material that is mostly obtained over a telephone (Hirson, French, & Howard, 1995: 230; Baltisberger & Hubbuch, 2010) – thus the frequency domain information available to listeners is reduced. First, the recognition of a foreign accent helps narrowing down a group of suspects in cases where an expert has to establish the identity of an individual based solely on his/her voice (speaker profiling: Ellis, 1994; French, 2007). Second, a number of governments use LADO (Linguistic Analysis for the Determination of Origin) to establish the geographical origin of an individual based solely on his/her voice, in cases where the claim of this individual to originate from a particular region is doubted (Baltisberger & Hubbuch, 2010). Foreign accent recognition could be a crucial part of the LADO analysis, since some individuals use L2 speech as a form of voice disguise during LADO interviews. In such cases, it is of particular interest to identify acoustic cues for the recognition of the subject's L1 (Priska Hubbuch, LINGUA – LADO section of the Swiss Federal Office for Migration, personal communication). However, research on foreign accent recognition with other L1/L2-pairs is needed before results can be applied in forensic casework.

## 6. Summary and conclusion

The present study investigated the recognition of French- and English-accented German L2 speech by Swiss German listeners, based on time domain characteristics. Different signal degradation procedures were applied to foreign-accented speech and subsequently used in a between-subject perception experiment. The type of temporal information contained in the delexicalised stimuli differed between the signal conditions: Noise vocoded speech is strongly degraded in the spectral domain and does not contain periodicity; segment durations are not or hardly perceivable. Subjects' attention is drawn to amplitude envelope temporal characteristics, or syllable beats. Monotone 1-bit requantised speech, on the other hand, lacks amplitude envelope information as well as f0 movements and most spectral cues, but allows the processing of segment durations and voice timing. Monotone *sasasa*-speech lacks all spectral information as well as original amplitude envelope and f0 movements; it contains cues to voice timing only.

The results reported in the present paper show that the time domain of speech is important for the recognition of foreign accents: Speech can be strongly degraded in the frequency domain and still provide enough cues for listeners to recognise a French or an English accent in German speech. We illustrated that different types of durational cues allow for higher or lower identification scores. Segment durations seem to be an L2 temporal characteristic that is (a) affected by interferences from the speaker's L1 and (b) perceptually salient to listeners. The same holds, to a lesser degree, for amplitude envelope durational cues. An additional finding of this study is that the stronger speech is degraded in the frequency domain, the more difficult it is for listeners to recognise foreign accents.

## Acknowledgements

This research was supported by the Swiss National Science Foundation (SNSF; grant number: 100015_135287). The authors would like to thank their subjects, speakers as well as listeners, for their contribution to this experiment. Furthermore, they thank Adrian Leemann and Stephan Schmid for extremely valuable feedback on a first version of this manuscript. Thanks to Stephan Schmid for the translation of the sentence material. Further thanks go to two anonymous reviewers and the associate editor, Ocke-Schwen Bohn, for their helpful comments on an earlier version of this manuscript.

## Appendix A. Reading materials

01 Die Frau des Apothekers weiss immer, was sie will.
02 Das Theater hat viele neue Aufführungen geplant.
03 Er wollte sich seiner Schwächen einfach nicht bewusst werden.
04 Der öffentliche Verkehr lässt viel zu wünschen übrig.
05 Die schlechte Zahlungsbilanz lässt mich nicht zur Ruhe kommen.
06 Die Eltern geben ihm keine finanzielle Unterstützung.
07 Der starke Frühlingsregen hat grossen Schaden angerichtet.
08 Der schnellste Zug ist immer noch der ICE.
09 Der Wiederaufbau der Stadt wird sehr lange dauern.
10 Das Bildungsministerium hat den einfachsten Weg gewählt.
11 Diese Konditorei macht ausgezeichnete Kuchen.
12 Dieses Geschäft bietet sehr preisgünstige Ware an.
13 Sie haben die Wahrheit erst entdeckt, als er auspackte.
14 Für meine Mannschaft wird der Sieg ein Kinderspiel sein.
15 Die Meinungsumfragen sagen einen Sieg der Rechten voraus.
16 Die Strassen der Innenstadt wurden von der Polizei gesperrt.
17 Ein berühmtes Bild wurde aus dem Kunsthaus gestohlen.
18 Der Müssiggang ist bekanntlich aller Laster Anfang.

## References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.
Adams, C., & Munro, R. R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude and duration in the connected utterance of some native and non-native speakers of English. *Phonetica*, *35*, 125–156.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*(*4*), 529–555.
Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(*2*), 351–373.
Auer, P. (2001). Silben- und akzentzählende Sprachen. In: M. Haspelmath, E. König, & W. Oesterreicher (Eds.), *Language typology and language universals. An international handbook. Sprachtypologie und sprachliche Universalien. Ein internationales Handbuch*, Vol. 2 (pp. 1391–1399). Berlin/New York: de Gruyter.
Baltisberger, E., & Hubbuch, P. (2010). LADO with specialized linguists – The development of LINGUA's working method. In: K. Zwaan, M. Verrips, & P. Muysken (Eds.), *Language and origin: The role of language in European asylum procedures* (pp. 9–19). Nijmegen: Wolf Legal Publishers.
Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica*, *65*, 131–147.
Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer*. Computer program. ⟨http://www.praat.org/⟩.
Boula de Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, *63*(*4*), 247–267.
Boula de Mareüil, P., Vieru-Dimulescu, B., Woehrling, C., & Adda-Decker, M. (2008). Accents étrangers et régionaux en français. *Traitement Automatique des Langues*, *49*(*3*), 135–163.
Bush, C. N. (1967). Some acoustic parameters of speech and their relationships to the perception of dialect differences. *TESOL Quarterly*, *1*(*3*), 20–30.
Carter, P. (2005). Quantifying rhythmic differences between Spanish, English, and Hispanic English. In: S. G. Randall, & J. R. Edward (Eds.), *Theoretical and experimental approaches to romance linguistics* (pp. 63–75). Amsterdam: Benjamins.
Council of Europe (2013). Common European framework of reference for languages: Learning, Teaching, Assessment. ⟨http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf⟩ (accessed 26.07.13).
Cummins, F., Li, Chenxia, & Wang, Bei (2013). Coupling among speakers during synchronous speaking in English and Mandarin. *Journal of Phonetics*, *41*, 432–441.
Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*, 145–171.
Cunningham-Andersson, U., & Engstrand, O. (1987). Perceived strength and identity of foreign accent in Swedish. *Phonetica*, *46*, 138–154.
Dauer, R. M. (1983). Stress-timing and syllable-timing reanalysed. *Journal of Phonetics*, *11*, 51–52.
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology*, *134*(*2*), 222–241.
Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC. In: P. Karnowski, & I. Szigeti (Eds.), *Language and language-processing. Proceedings of the 38th linguistics colloquium 2003, Piliscsaba, Hungary* (pp. 231–241). Lang: Frankfurt am Main etc.
Dellwo, V. (2008). The role of speech rate in perceiving speech rhythm. *Proceedings of the 4th International Conference on Speech Prosody 2008, Campinas, Brazil* (pp. 375–378).
Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence* (Ph.D. thesis). Bonn: University of Bonn.
Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In: *Proceedings of the 16th international congress of phonetic sciences (ICPhS) 2007*, *Saarbrücken*, *Germany* (pp. 1129–1132).
Dellwo, V., Leemann, A., & Kolly, M. -J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In: *Proceedings of interspeech 2012, Portland, USA*.
den Os, E. (1988). *Rhythm and tempo of Dutch and Italian*. Utrecht: Elinkwijk.
Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility. Evidence from four L1s. *Studies in Second Language Acquisition*, *19*(*1*), 1–16.
Donaldson, W. (1992). Measuring recognition memory. *Journal of Experimental Psychology: General*, *121*(*3*), 275–277.
Ellis, S. (1994). The Yorkshire Ripper enquiry: Part 1. *Forensic Linguistics*, *1*, 197–206.
Ferragne, E., & Pellegrino, F. (2004). Rhythm in read British English: Interdialect variability. In: *Proceedings of the 8th international conference on spoken language processing 2004, Jeju, Korea* (pp. 1573–1576).
Fourcin, A., & Dellwo, V. (2009). *Rhythmic classification of languages based on voice timing*. London: UCL Eprints. ⟨http://eprints.ucl.ac.uk/15122/⟩ (accessed 26.07.13).
French, P. (2007). Caller on the line: An illustrated introduction to the work of a forensic speech scientist. *Medico-Legal Journal*, *75*(*3*), 83–96.
Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In: C. Gussenhoven, & N. Warner (Eds.), *Laboratory phonology*, *7* (pp. 515–545). Berlin/New York: Mouton de Gruyter.
Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.
Guntern, M. (2011). Erkennen von Dialekten anhand von gesprochenem Schweizerhochdeutsch. *Zeitschrift für Dialektologie und Linguistik*, *78*(*2*), 155–187.
Gutiérrez Díez, F., Dellwo, V., Gavaldà, N., & Rosen, S. (2008). The development of measurable speech rhythm during second language acquisition. *Journal of the Acoustical Society of America*, *123*(*5*), 3886.
Hirson, A., French, P., & Howard, D. (1995). Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. In: J. Windsor Lewis (Ed.), *Studies in general and english phonetics in honour of Professor J.D. O'Connor* (pp. 230–240). London: Routledge.
Holm, S. (2008). *Intonational and durational contributions to the perception of foreign-accented Norwegian. An experimental phonetic investigation* (Ph.D. thesis). Norwegian University of Science and Technology.
Jilka, M. (2000). *The contribution of intonation to the perception of foreign accent* (Ph.D. thesis). Stuttgart: University of Stuttgart.
Kohler, K. (2009). Rhythm in speech and language. A new research paradigm. *Phonetica*, *66*(*1–2*), 29–45.
Kolly, M.-J. (2013). Akzent auf die Standardsprachen: Regionale Spuren in 'Français Fédéral' und 'Schweizerhochdeutsch'. *Linguistik online*, *58*(*1*), 37–76.
Kolly, M.-J., & Dellwo, V. (2013). (How) do listeners perceive the origin of a foreign accent? *Travaux neuchâtelois de linguistique*, *59*, 127–148.
Koster, C. J., & Koet, T. (1993). The evaluation of accent in the English of Dutchmen. *Language Learning*, *43*(*1*), 69–92.
Kumpf, K., & King, R. W. (1997). Foreign speaker accent classification using phoneme-dependent accent discrimination models and comparisons with human perception benchmarks. In: *Proceedings of the 5th European conference on speech communication and technology, 1997, Rhodes (Greece)* (pp. 2323–2326).
Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition*, *93*, 225–254.
Leemann, A. (2011). Einfluss der Schweizerdeutschen Phonologie auf die Stimmhaftigkeit von Frikativen im L2-Englischen. Poster presented at the *7th conference* '*Phonetik und Phonologie*' *2011, Osnabrück, Germany*.
Leemann, A., & Siebenhaar, B. (2008). Perception of dialectal prosody. In: *Proceedings of interspeech 2008, Brisbane, Australia* (pp. 524–527).
Leemann, A., Dellwo, V., Kolly, M. -J., & Schmid, S. (2012). Rhythmic variability in Swiss German dialects. In: *Proceedings of the 6th international conference on speech prosody 2012, Shanghai, PRC* (pp. 607–610).
Licklider, J. C. R., & Pollack, I. (1948). Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech. *Journal of the Acoustical Society of America*, *20*(*1*), 42–51.
Lippi-Green, R. (1997). *English with an accent: Language, ideology and discrimination in the United States*. London/New York: Routledge.
Lloyd James, A. (1929). *Historical introduction to French phonetics*. London: ULP.
Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*, 381–400.
Munro, M. J. (1995). Nonsegmental factors in foreign accent. Ratings of filtered speech. *Studies in Second Language Acquisition*, *17*(*1*), 17–34.
Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. The role of speaking rate. *Studies in Second Language Acquisition*, *23*(*4*), 451–468.
Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*, 626–637.
Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(*3*), 756–766.
Neuhauser, S. (2011). Foreign accent imitation and variation of VOT and voicing in plosives. In: *Proceedings of the 15th international congress of phonetic sciences (ICPhS) 2003, Barcelona, Spain* (pp. 1462–1465).
Park, H. (2013). Detecting foreign accent in monosyllables: The role of L1 phonotactics. *Journal of Phonetics*, *41*, 78–87.
Pike, K. (1945). *The intonation of American English*. Ann Arbor: University of Michigan Press.
Quené, H, & van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*, 911–918.
R Core Team (2013). *R. A language and environment for statistical computing*. Version 3.0.1. Vienna. ⟨http://www.R-project.org⟩.
Ramus, F., Dupoux, E., & Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In: *Proceedings of the 15th international congress of phonetic sciences (ICPhS) 2003, Barcelona, Spain* (pp. 1–6).
Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, *105/1*, 512–521.
Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*, 265–292.
Schmid, S. (2012a). The pronunciation of voiced obstruents in L2 French: A preliminary study of Swiss German learners. *Poznań Studies in Contemporary Linguistics*, *48*(4), 627–659.
Schmid, S. (2012b). Phonological typology, rhythm types and the phonetics-phonology interface. A methodological overview and three case studies on Italo-Romance dialects. In: A. Ender, A. Leemann, & B. Wälchli (Eds.), *Methods in contemporary linguistics. A Festschrift in honour of Iwar Werlen* (pp. 45–68). Berlin/New York: Mouton de Gruyter.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*, 303–304.
Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics*, *25*, 1–24.
Taylor, D. S. (1981). Nonnative speakers and the rhythm of English. *International Review of Applied Linguistics in Language Teaching*, *19*, 221–226.
Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America*, *134*, 628–639.
Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*, *124*(*2*), 34–39.
Trouvain, J., & Gut, U. (Eds.). (2007). *Non-native prosody. Phonetic description and teaching practice*. Berlin/New York: de Gruyter.
Van Els, T., & De Bot, K. (1987). The role of intonation in foreign accent. *Modern Language Journal*, *71*(*2*), 147–155.
White, L., & Mattys, S. L. (2007a). Calibrating rhythm: First and second language studies. *Journal of Phonetics*, *35*, 501–522.
White, L., & Mattys, S. L. (2007b). Rhythmic typology and variation in first and second languages. In: P. Prieto, J. Mascaró, & M.-J. Solé (Eds.), *Segmental and prosodic issues in romance phonology* (pp. 237–257). Amsterdam/Philadelphia: Benjamins.
White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, *66*(*4*), 665–679.
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America*, *127*, 1559–1569.
Winters, S., & O'Brien, M. G. (2013). Perceived accentedness and intelligibility. The relative contributions of f0 and duration. *Speech Communication*, *55*, 486–507.

# 5

# Foreign accent recognition based on temporal information contained in lowpass-filtered speech

This chapter contains a reprint of the paper: Kolly, M.-J., Leemann, A., Dellwo, V. (2014). Foreign accent recognition based on temporal information contained in lowpass-filtered speech. *Proceedings of Interspeech 2014*, Singapore: 2175–2179.

Chapter 4 demonstrated that listeners can identify speaker origin in French- and English-accented German based on primarily temporal information. However, when frequency domain information was completely absent from stimuli, accent identification performance was at chance. One of the hypotheses to arise from the conclusions in Chapter 4 is that the signal types used for the stimuli may have hampered listeners' accent identification performance, since they were unlikely to occur in in natural environments. We thus considered the possibility that listeners could identify foreign accents based on time domain features if such features were presented in a signal type they were familiar with. For the experiments presented in this paper, we used lowpass-filtered stimuli. These were assumed to sound relatively familiar to listeners, as listeners are used to hearing speech through walls or closed doors. Stimuli were created with the materials from Chapter 4 and they were, again, heavily degraded in the frequency domain:

- ▷ Lowpass-filtered speech with a cutoff frequency of 300 Hz contained information on fundamental frequency variability as well as amplitude envelope and voicing temporal features.
- ▷ Monotonized lowpass-filtered speech with a cutoff frequency of 300 Hz contained primarily temporal cues related to the pulsing of the amplitude envelope and to voicing temporal patterns.

The outcome of this research can be outlined as follows:

⇒ Both signal types allowed listeners to identify foreign accents above chance.

⇒ Fundamental frequency variability information seems to facilitate accent identification: monotonized lowpass-filtered speech was identified with lower performance.

⇒ French-accented German was identified with higher performance than English-accented German only in monotonized lowpass-filtered speech.

⇒ There was an effect of speaker on listeners' accent identification performance, which was only moderately correlated with speakers' foreign accent strength.

When comparing the result of monotonized lowpass-filtered speech with that of monotonized *sasasa*-speech (which did not allow for accent identification above chance, see Chapter 4), we note that the former generated higher accent identification performance. Compared to *sasasa*-speech, lowpass-filtered speech contains similar information on voicing temporal patterns. However, it additionally contains information on intensity timing, and lacks the segmental [s] and [a] information. Given these results, we concluded that the combination of voicing and intensity temporal cues in lowpass-filtered speech may yield higher accent identification performance than voicing temporal cues on their own. On the other hand, it may be listeners' familiarity with lowpass-filtered speech that facilitated the processing of temporal cues to foreign accent in this signal. Furthermore, the synthetic [s] and [a] segmental information may divert listeners' attention in *sasasa*-speech.

Our choice of a cutoff frequency of 300 Hz was motivated by the possibility of including fundamental frequency information but excluding segmental information, particularly cues to vowel quality, from these signals (see Section 1 of the present paper). However, certain frequency domain cues to the quality of consonants and even vowels may have remained in the stimuli and facilitated listeners' task. We therefore conducted a further experiment that presented listeners with time domain information alone, in stimuli that were as close to natural speech as possible. This experiment is presented in Chapter 6.

As the effect of speaker on accent identification performance was only moderately correlated with speakers' foreign accent strength, we hypothesize that speakers employ different strategies in the time domain when producing non-native speech, regardless of their proficiency. Such speaker-individual temporal patterns of non-native speech are explored in Chapter 7.

# Foreign accent recognition based on temporal information contained in lowpass-filtered speech

*Marie-José Kolly*[1], *Adrian Leemann*[1], *Volker Dellwo*[1]

[1]Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Switzerland
{marie-jose.kolly, adrian.leemann}@pholab.uzh.ch, volker.dellwo@uzh.ch

## Abstract

Can the foreign accent of a speaker be recognized based on suprasegmental temporal information? For a perception experiment we created stimuli based on German sentences read by six French and six English speakers. These foreign-accented sentences were manipulated by (1) applying a lowpass filter with a cutoff frequency of 300 Hz and (2) applying the same lowpass filter and monotonizing F0. In a between-subject 2AFC perception experiment we tested the accent recognition ability of 15 Swiss German listeners per signal manipulation condition. The results showed that speakers' native language could be recognized above chance in both conditions. However, listeners obtained significantly lower recognition scores in the monotonized condition. Furthermore, higher recognition scores were obtained for French-accented speech in the monotonized condition, a result that is discussed in light of research on speech rhythm. We further report an effect for *speaker* within each accent group. The results suggest that suprasegmental temporal information allows for foreign accent recognition to some degree.

**Index Terms**: foreign accent recognition, speaker origin, lowpass-filtered speech, temporal characteristics, speech rhythm

## 1. Introduction

"Judging by your accent, you must be French" – people readily engage in foreign accent recognition tasks when listening to second language speech. But how, i.e. based on which cues, do listeners make decisions on a speaker' native language (L1)? Second language (L2) speech differs from native speech in a number of characteristics, and some of these characteristics are perceptually salient to listeners. For example, /r/ in the English word *foreign* is typically pronounced as a uvular trill [ʀ] or fricative [ʁ] by French speakers and as an alveolar trill [r] by Italian speakers. Provided that an English listener has common knowledge of French, an [ʀ] in *foreign* – among other cues – may lead him/her to guess the speaker's L1 as being French. Research has shown the importance of segmental cues for foreign accent recognition [1, 2].

The importance of suprasegmental cues for foreign accent recognition has been investigated by a handful of studies, which focused on frequency domain information. [3], for example, found that the absence of segmental accent-cues still allows listeners to recognize speaker origin in L2 speech, based on cues to f0 variability (i.e., intonation) and segment durations. [4] also demonstrated the importance of cues to f0 variability for foreign accent recognition. [4] further found that listeners were no longer able to recognize foreign accents in lowpass-filtered speech below 350 Hz, which suggested that time domain cues alone are not sufficient for this type of task. However, the multiple choice listening task used in [4] allowed the response "I don't know", an option that was frequently chosen by listeners. An alternative forced choice (AFC) experiment design may have yielded different results. Moreover, evidence from the field of dialect recognition suggested that time domain characteristics allow for dialect recognition: In a 4AFC experiment, listeners were able to recognize 3/4 Swiss German dialects in lowpass-filtered speech below 250 Hz [5].

The contribution of suprasegmental time domain information to foreign accent recognition was shown in [6]: Listeners were able to recognize foreign accents based on primarily temporal cues contained in 1-bit requantized speech [6, 7], for which the bit-rate of the acoustic signal was reduced to 1-bit, and in 6-band noise vocoded speech [8], for which amplitude envelopes were extracted from 6 frequency bands and used to modulate white noise. The latter sounds like a harsh whisper [9]. However, in signal manipulation conditions where listeners had no access to cues from the frequency domain (e.g. in 3-band noise vocoded speech, or in monotonized *sasasa*-speech [10]), foreign accent recognition was no longer possible [6]. The outcome of this research suggested that either it was the interplay between time and frequency domain characteristics that enabled foreign accent recognition, or that time-domain-only signal conditions that occur in natural situations would possibly yield different results and enable foreign accent recognition. In fact, 3-band noise vocoded speech and *sasasa*-speech are extremely distorted speech signals: In 3-band noise vocoded speech, the source signal of speech is replaced with white noise. In our *sasasa*-speech, every voiced speech interval was replaced with the same [a]-sound and every unvoiced speech interval with the same [s]-sound. Such "speech"-signals are unlikely to occur in everyday situations, which was mirrored by listeners' feedback in [6].

The present contribution, a follow-up experiment on [6], explores foreign accent recognition based on time domain characteristics contained in lowpass-filtered speech. This type of signal may appear more natural for listeners, since lowpass-filtered speech occurs in everyday situations: When a conversation is heard through a closed door, for example, or through a thick wall [11]. In this kind of situation, a listener may try to guess the language, accent or identity of the speakers. These guesses are confirmed once the speakers open the door: Their language, accent or identity becomes apparent to the listener. Listeners are therefore assumed to be familiar with the correspondence between unfiltered and filtered speech (e.g. of a particular language, accent, or speaker).

We aimed at using stimuli that contain no information on speech segmental content in order to isolate suprasegmental temporal and rhythmic features. We therefore filtered speech with a cutoff frequency of 300 Hz. We did not use a higher cutoff, since we wanted to exclude cues to vowel qualities: F1-values of vowels below 300 Hz are rather unusual in French, English and German [12, 13, 14]. We did not use a lower

cutoff, since female mean f0-values often attain 250 Hz in read speech [12, 15] and we wanted to include cues to f0 variability in one of our signal manipulation conditions – henceforth *lowpass* condition. We used the same filter for our second signal manipulation condition, and additionally monotonized f0 – henceforth *lowpass.monotonized* condition.

## 2. Materials and methods

### 2.1. Subjects

In a between-subject design, we tested a total of 30 Swiss German listeners: 15 listeners were tested with the *lowpass* condition (6 male / 9 female) and 15 with the *lowpass.monotonized* condition (5 male / 10 female). Subjects were university students and aged between 18 and 31 (M=23, SD=3). None of the subjects reported significant problems with hearing or sight. Their school education in second language French and English was comparable: French is usually introduced as a second, English as a third language in Swiss German schools. Swiss German university students have studied French and English for approximately 11 and 6 years respectively. We assumed listeners to have a similar level of familiarity with French and English speakers of German respectively, since the listener group was homogenous in terms of age and educational level.

### 2.2. Material

Stimuli were created based on speech from 6 French and 6 English native speakers (3 males / 3 females each). French speakers' self-assessed proficiency in German was intermediate (B1 to B2), English speakers' proficiency ranged from beginner to intermediate (A1 to B2), cf. [16]. Speakers' foreign accent degree was rated on a 5-point scale (1=very strong accent, 5=no accent) by 10 Swiss German listeners in a previous experiment. Results revealed that the degree of accentedness did not differ between the French and the English speakers [6].

Speakers read a list of 18 sentences that contained 12–16 syllables each. They were recorded with a *Fostex FR-2LE* solid-state recorder and a *Sennheiser MKE 2p-c* clip-on microphone (48 kHz, 16 bit) in a quiet room at the University of Zurich or in their own homes. We selected different sets of 9 sentences per speaker such that the experiment contained 108 sentences. Every sentence appeared 6 times in the experiment: 3 times with a French and 3 times with an English accent (cf. [6]).

Lowpass-filtered stimuli were constructed using Praat [17]. Every sentence was lowpass-filtered with a cutoff frequency of 300 Hz and a smoothing-value of 50 Hz (width between pass and stop, cf. [17]). An example of natural and lowpass-filtered speech from our stimuli is shown in Figure 1. To create monotonized lowpass-filtered speech, we first removed octave jumps automatically [17] and then replaced the pitch points of every sentence with the mean pitch value of the sentence. We used this procedure since averaging all male and all female sentences to a specific f0 mean produced stimuli that sounded unnatural (as judged by informal listening tests). We ran t-tests to examine the effect of the factor *accent* on mean f0: We did not find significant differences in f0 means between the French and the English accent group, neither for the *lowpass* (t=-0.53, ns, df=106) nor for the *lowpass.monotonized* condition (t=-0.08, ns, df=106). The accent recognition scores reported in section 3 are thus assumed to be independent from speakers' mean f0s. Finally, every stimulus-sentence was scaled to an intensity of 75 dB.

Informal perception experiments showed that listeners could not retrieve frequency domain information other than f0 variability from our stimuli. We presented listeners with two filtered vowel sounds and two categories as options: They were asked to decide which category belonged to which sound. Listeners were not able to identify single vowels in our lowpass-filtered speech. We understand this as evidence that our stimuli did not contain sufficient frequency domain cues that may have enabled the identification of individual vowel segments. Since frequency domain cues to consonants lie higher than 300 Hz this also means that consonantal distinctions could not be performed based on spectral envelope characteristics of consonants. In summary it can be said that our lowpass-filtered speech predominantly contained cues to voicing characteristics, i.e. on- and offset of voice as well as – in the *lowpass* condition – to changes of f0 over time (i.e., intonation).

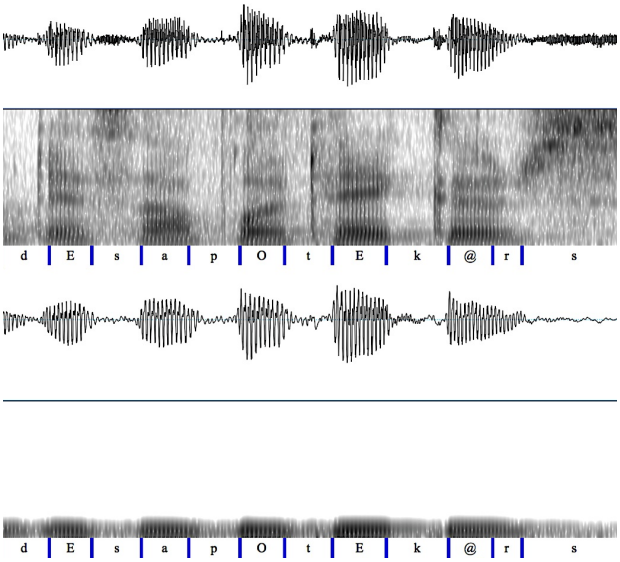



Figure 1: *SAMPA-transcribed waveform and spectrogram of the phrase* des Apothekers *'of the pharmacist' spoken by a native English speaker; natural (top) and lowpass-filtered (bottom) speech.*

### 2.3. Procedure

Listeners were tested in a quiet room at the University of Zurich. The experiment lasted between 15 and 25 mins. Subjects heard the stimuli through high-quality earphones, where the order of the stimuli was randomized separately for each subject. For both signal conditions, the sentence corresponding to the acoustic stimulus was presented on a laptop screen two seconds preceding the acoustic stimulus and during acoustic stimulus presentation. Following the presentation of each stimulus, subjects had to decide whether they heard French- or English-accented German by clicking on the corresponding button on a laptop computer, using the experiment interface shown in Figure 2. They further indicated the confidence of their response on a 3-point scale (1 = sure, 2 = rather unsure, 3 = only guessing).


Figure 2: *Experiment interface; to give their response, listeners clicked on one of the small blue rectangles.*

### 2.4. Data analysis and statistics

Based on each listener's responses we calculated *d'*, a measure derived from signal detection theory, based on the numbers of hits, false alarms, correct rejections and misses [18]. *d'* is obtained from each listeners' hit rate and false alarm rate: *d'* = *z-value*(hit)–*z-value*(false alarm). It measures listeners' sensitivity, i.e. their ability to discriminate two types of signals – French- vs. English-accented German – while canceling out response bias. Perfect sensitivity is reached at a *d'*-value of 4, whereas a *d'*-value of 0 indicates sensitivity at chance level. Normality of the *d'*-distribution was checked by visual inspection of quantile plots. To obtain listeners' recognition scores for each of the two signal types – i.e. accents – separately, we calculated the percentage of listeners' correct responses: *%correct* = (hits + correct rejections)/(hits + false alarms + correct rejections + misses). Statistical analyses were conducted using *R* [19]. We used two-sided t-tests and tested at a significance level of α=0.05.

## 3. Results

Results are presented as follows: In 3.1 we report the findings on listeners' general ability to recognize French- and English-accented German in *lowpass* and *lowpass.monotonized* speech. 3.2 shows the effect of signal manipulation condition on recognition performance. 3.3 presents results on listeners' recognition performance for French- and English-accented speech separately. 3.4 shows the effect of speaker on listeners' recognition performance.

### 3.1. Listeners' accent recognition performance

T-tests showed that listeners were able to recognize French- and English-accented German above chance in the *lowpass* condition (t=7.15, p<0.0001, df=14) as well as in the *lowpass.monotonized* condition (t=6.09, p<0.0001, df=14). This result is presented in Figure 3. Compared to *d'*-values of 4 for perfect sensitivity, the values reported here are fairly low (*lowpass*: M=0.61, *lowpass.monotonized*: M=0.39). However, this is in line with other investigations that use strongly degraded speech: [20], for example, report mean *d'*-values of 0.17 and 0.30 for listeners' recognition of English dialects in monotonized *sasasa*-speech.

### 3.2. Effect of signal manipulation condition

As can be seen in Figure 3, the two boxplots' interquartile ranges only overlap to a small degree: There was a significant difference between the signal conditions (t=2.04, p=0.05, df=26), with listeners obtaining higher accent recognition scores in *lowpass* (blue; M=0.61, SD=0.33) than in *lowpass.monotonized* (red; M=0.39, SD=0.25). Furthermore, the signal conditions differed significantly in listeners' certainty of response (t=-10.43, p<0.0001, df=3201), with listeners reporting higher degrees of certainty when making decisions about accents in *lowpass* (M=1.60, SD=0.68) as opposed to *lowpass.monotonized* (M=1.87, SD=0.76).

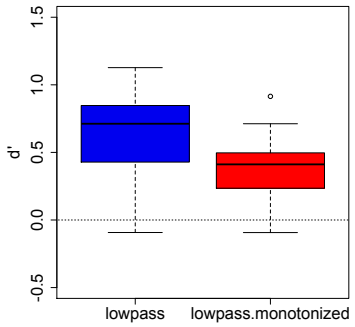



Figure 3: *Boxplots of* d' *for two signal conditions; dotted line = performance at chance.*

### 3.3. Effect of accent type

We calculated the percentage of correct responses for the French- and the English-accented stimuli separately: *%correct*. T-tests showed that French-accented German obtained higher recognition scores in *lowpass.monotonized* speech (red; t=-2.08, p<0.05, df=28) but not in *lowpass* speech (blue; t=-0.24, ns, df=28). This result is presented in Figure 4.

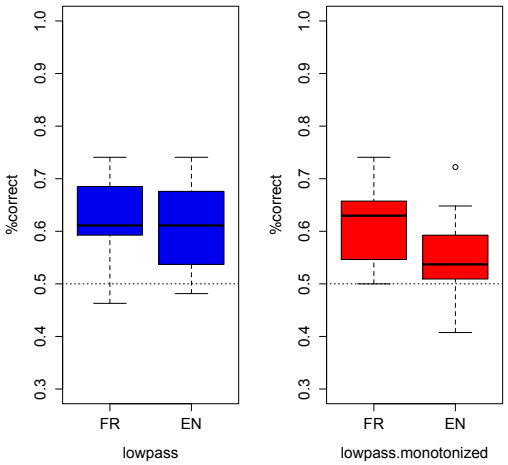



Figure 4: Boxplots of %correct for two accents by signal condition; dotted line = performance at chance.

### 3.4. Effect of speaker

A univariate ANOVA with *%correct* as the dependent variable shows that listeners' recognition scores differed depending on the speaker who articulated the sentences, within the French (F(5, 24)=9.04, p<0.01) as well as within the English (F(5, 24)=9.35, p<0.01) accent group (signal manipulation conditions pooled). This result is illustrated in Figure 5. We found a moderate correlation between *%correct* for each

speaker and speakers' accent degree (r=-0.43; French and English speakers pooled).

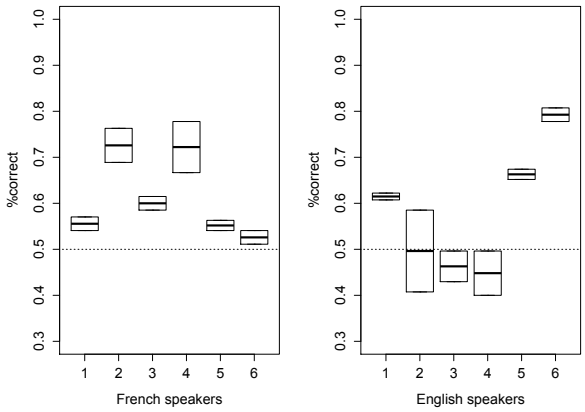



Figure 5: Boxplots of %correct for six French (left) and six English speakers (right); dotted line = performance at chance.

## 4. Discussion & Conclusion

Our results showed that listeners were able to recognize French- and English-accented speech above chance in the *lowpass* as well as in the *lowpass.monotonized* condition.

The result of the *lowpass* condition suggests that a 2AFC task allows listeners to recognize foreign accents when only cues to time domain and to f0 variability are available – which was not possible in [4], where listeners had the possibility to respond "I don't know". Our findings reflect similar results as research on dialect recognition in lowpass-filtered speech below 250 Hz [5], and findings on language discrimination by newborns in lowpass-filtered speech below 400 Hz [21] or by adults in lowpass-filtered speech below 180 Hz [22].

Our data on *lowpass.monotonized* speech shows that listeners are able to recognize foreign accents when no frequency domain information is present. Similar recognition performances were observed in 6-band noise-vocoded speech, a signal condition that allows listeners to access frequency domain information to some degree [6]. However, [6] showed that listeners performed at chance for signal manipulation conditions that did not contain frequency domain information, in particular for monotonized *sasasa*-speech based on voiced and voiceless intervals (see section 1), which contains similar suprasegmental temporal information as our *lowpass.monotonized* stimuli: Lowpass-filtered speech contains information about voice timing and information about intensity timing. *Sasasa*-speech contains information about voice timing only. Two explanations can be put forth for the discrepancy in listeners' accent recognition performance in these two signal conditions. (1) *Lowpass.monotonized* speech contains cues to intensity, which was not the case for the monotonized *sasasa*-speech used in [6]. The combination of time domain and intensity domain cues may have been important for listeners' ability to recognize foreign accents when no frequency domain information was available. (2) *Sasasa*-speech is unlikely to occur in natural situations; however, listeners can be assumed to be familiar with lowpass-filtered speech (cf. section 1), which may affect their recognition performance.

Our results further showed that listener performance and confidence differed significantly with regard to the signal manipulation condition: accent recognition performance as well as confidence was higher in *lowpass* than in *lowpass.monotonized* speech. It is plausible that this has to do with the fact that *lowpass* stimuli are signal-degraded to a lesser extent than *lowpass.monotonized* stimuli, i.e. they contain more cues – frequency domain cues in particular – that listeners can use to solve the accent recognition task. From this we infer that the absence of intonation in the *lowpass.monotonized* condition affected listener performance, but still allowed for accent recognition above chance. Similarly, [4] and [6] showed that listeners' accent recognition performance decreases as frequency domain information is reduced in signal-degraded speech.

We found that listeners' performance was significantly higher for the French-accented than for the English-accented stimuli in the *lowpass.monotonized*, but not in the *lowpass* condition. This suggests that French-accented German sounds perceptually more salient in the suprasegmental temporal domain than English-accented German. If interferences from speakers' L1 account for this, one may speculate that English is in fact closer to German than French in its suprasegmental temporal features – as it has been suggested by the literature on speech rhythm, which classified languages in rather "syllable-timed" (e.g. French) and rather "stress-timed" (e.g. English, German) [23–26], or in more and less "regular" [27].

We further found a significant effect of speaker on listeners' accent recognition performance, for the French- as well as for the English-accented stimuli. However, listeners' recognition performance for each speaker was only moderately correlated with speakers' accent degree. Since accent degree was rated based on natural speech (cf. [6]) it may be that listeners focused on different cues when listening to filtered speech, as reported in [28] – where it was found that listeners' ratings of foreign accent degree in natural speech and in filtered speech were not correlated. However, more research is needed before any conclusions can be drawn on our data.

Implications of this research can be found in the domain of second language acquisition: Our results suggest that suprasegmental temporal features are especially salient in French speakers' German speech. If an alleviation of foreign accentedness is desired, then learners of a second language that differs from their native language in its suprasegmental temporal organization may practice this type of feature in particular. From a more practical viewpoint, it has been shown that temporal features of foreign-accented speech have an effect on speakers' intelligibility [29].

Conducting research with more forensic phonetic applications in mind, we plan further perception experiments with our *lowpass* and *lowpass.monotonized* conditions without presenting visual information on sentence content. This task will be more similar to forensically relevant situations. For example, an ear-witness may hear a crime-related conversation through a closed door, and subsequently be asked to describe the linguistic profiles of the speakers s/he heard.

## 5. Acknowledgements

This research was supported by the Swiss National Science Foundation (SNSF; grant number: 100015_135287) and a grant of the Department of General Linguistics, University of Zurich. The authors would like to thank their subjects for their contribution to this experiment. We also thank Stephan Schmid for his expert advice on foreign-accented speech.

# 6. References


[1] Cunningham-Andersson, U. and Engstrand, O., "Perceived strength and identity of foreign accent in Swedish", Phonetica, 46:138-154, 1987.

[2] Boula de Mareüil, P., Vieru-Dimulescu, B., Woehrling, C. and Adda-Decker, M., "Accents étrangers et régionaux en français", Traitement Automatique des Langues, 49:135-163, 2008.

[3] Boula de Mareüil, P. and Vieru-Dimulescu, B., "The contribution of prosody to the perception of foreign accent", Phonetica, 63:247-267, 2006.

[4] Van Els, T. and De Bot, K., "The role of intonation in foreign accent", Modern Language Journal, 71:147-155, 1987.

[5] Leemann, A. and Siebenhaar, B., "Perception of dialectal prosody", Proceedings of Interspeech 2008, Brisbane, Australia: 524-527, 2008.

[6] Kolly, M.-J. and Dellwo, V., "Cues to linguistic origin: The contribution of speech temporal information to foreign accent recognition", Journal of Phonetics, 42:12-23, 2014.

[7] Licklider, J.C.R. and Pollack, I., "Effects of differentiation, integration, and infinite peak clipping upon the intelligibility of speech", Journal of the Acoustical Society of America, 20:42-51, 1948.

[8] Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J. and Ekelid, M., "Speech recognition with primarily temporal cues", Science, 270:303-304, 1995.

[9] Scott, S.K., "The neurobiology of speech perception", in A. Cutler [Ed], Twenty-first century psycholinguistics: Four cornerstones, 141-156, Mahwah, NJ: Erlbaum, 2005.

[10] Ramus, F. and Mehler, J., "Language identification with suprasegmental cues: a study based on speech resynthesis", Journal of the Acoustical Society of America, 105:512-521, 1999.

[11] Hervais-Adelman, A.G., Davis, M.H., Johnsrude, I.S., Taylor, K.J. and Carylon, R.P., "Generalization of perceptual learning of vocoded speech", Journal of Experimental Psychology: Human Perception and Performance, 37:283-295, 2011.

[12] Peterson, G.E. and Barney, H.L., "Control methods used in a study of the vowels", Journal of the Acoustical Society of America, 24:175-184, 1952.

[13] Hillenbrand, J., Getty, L.A., Clark, M.J. and Wheeler, K., "Acoustic characteristics of Amercian English vowels", Journal of the Acoustical Society of America, 97:3099-3111, 1995.

[14] Delattre, P., Comparing the phonetic features of English, German, Spanish and French, Heidelberg: Julius Groos, 1965.

[15] Künzel, H.J., Masthoff, H.R. and Köster, J.P., "The relation between speech tempo, loudness, and fundamental frequency: an important issue in forensic speaker recognition", Science and Justice, 35:291-295, 1995.

[16] Council of Europe, Common European framework of reference for languages: learning, teaching, assessment, http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf, accessed on 12 Jun 2014.

[17] Boersma, P. and Weenink, D., Praat: doing phonetics by computer, www.praat.org, 2013.

[18] Green, D.M, and Swets, J.A., Signal detection theory and psychophysics, New York: Wiley, 1966.

[19] R Core Team, R: A language and environment for statistical computing, version 3.0.1, http://www.R-project.org, 2013.

[20] White, L., Mattys, S.L. and Wiget, L., "Language categorization by adults is based on sensitivity to durational cues, not rhythm class", Journal of Memory and Language, 66:665-679, 2012.

[21] Nazzi, T., Bertoncini, J. and Mehler, J., "Language discrimination by newborns: toward an understanding of the role of rhythm", Journal of Experimental Psychology: Human Perception and Performance, 24:756-766, 1998.

[22] den Os, E., Rhythm and tempo of Dutch and Italian, Utrecht: Elinkwijk, 1988.

[23] Abercrombie, D., Elements of general phonetics, Edinburgh: Edinburth University Press, 1967.

[24] Pike, K., The intonation of American English, Ann Arbor: University of Michigan Press, 1945.

[25] Ramus, F., Nespor, M. and Mehler, J., "Correlates of linguistic rhythm in the speech signal", Cognition, 73:265-292, 1999.

[26] Grabe, E. and Low, E.L., "Durational variability in speech and the Rhythm Class Hypothesis", in C. Gussenhoven and N. Warner [Eds], Laboratory Phonology 7, 515-545, Berlin/New York: Mouton de Gruyter, 2002.

[27] Dellwo, V., "The role of speech rate in perceiving speech rhythm", Proceedings of Speech Prosody 2008, Campinas, Brazil: 275-278, 2008.

[28] Munro, M., "Nonsegmental factors in foreign accent: ratings of filtered speech", Studies in Second Language Acquisition, 17:17-34, 1995.

[29] Tajima, K., Port, R. and Dalby, J., "Effects of temporal correction on intelligibility of foreign-accented English", Journal of Phonetics, 25:1-24, 1997.

# 6

# Listeners use temporal information to identify French- and English-accented speech

This chapter contains a reprint of the paper: Kolly, M.-J., Boula de Mareüil, P., Leemann, A., Dellwo, V. (2017). Listeners use temporal information to identify French- and English-accented speech. *Speech Communication*, 86: 121–134.[1]

In Chapter 4 we suggested that time domain information may be of particular use to listeners in situations where frequency domain information is heavily degraded. If this is true, we would expect listeners' accent identification performance to show an additive effect when time and frequency domain cues are present in stimuli, as opposed to stimuli that contain time domain cues or frequency domain cues alone. Chapter 4 revealed that 6-band noise vocoded speech allows listeners to identify accents above chance; this signal condition contains time domain features as well as strongly reduced frequency domain features. These features can be separated as described below, in order to test listeners' performance when only one cue is presented.

The experiments presented in this chapter were conducted, on the one hand, to investigate whether each cue on its own still allows for accent identification and whether the combined presence of temporal and spectral cues boosts accent identification performance in 6-band noise vocoded speech. On the other hand, we suggested in Chapter 5 that monotonized lowpass-filtered speech below 300 Hz — albeit containing primarily temporal characteristics — may still contain reduced cues to specific vowels or consonants. This means that the question of whether listeners can identify foreign accents based on time domain characteristics alone in natural-sounding stimuli remains to be answered.

We therefore created stimuli that would sound as familiar as possible to listeners but contain time domain characteristics alone. These stimuli mainly contained segment durations, as the experiments from Chapter 4 have shown that this temporal feature may be salient to listeners. We further created stimuli that contain the strongly degraded frequency domain characteristics from 6-band noise vocoded speech alone. Stimuli were

[1]DOI: http://dx.doi.org/10.1016/j.specom.2016.11.006.

created using the 'prosody transplantation' method:

- ▷ For duration-transplanted native speech we took native German segments and modified their durations with segment durations of French- and English-accented German; stimuli contained only temporal features of the non-native speech.
- ▷ For duration-transplanted and 6-band noise vocoded non-native speech we modified the durations of French- and English-accented segments with native German segment durations and 6-band noise vocoded these signals; stimuli contained only the reduced frequency domain features of the non-native speech.

The principal findings reported in this paper are the following:

- ⇒ Both signal types, i.e., each cue on its own, allowed listeners to identify foreign accents above chance.
- ⇒ Frequency domain information, though severely degraded, yielded higher accent identification performance than time domain information.
- ⇒ An additive trend was observed when comparing these results to the experiment with 6-band noise vocoded speech where time and frequency domain cues are combined (see Chapter 4).
- ⇒ Listeners were biased towards perceiving French-accented German when stimuli featured uvular /r/s and towards perceiving English-accented German when they featured vocalized /r/s or when they lacked /r/.
- ⇒ We observed an effect of speaker on listeners' accent identification performance; this effect was not correlated with speakers' foreign accent strength.

We concluded that listeners make use of time domain information to identify speaker origin in non-native speech, but frequency domain information proved to be more salient in terms of foreign accent. The reported additivity of cues seems to confirm the hypothesis put forward in Chapter 4: time domain features facilitate, to some extent, the processing of rudimentary frequency domain information. Furthermore, we suggested that the method used to create our stimuli, i.e., the transplantation of non-native durations to native speech, has pitfalls: listeners took native German segments (i.e., the pronunciation of /r/) as a cue in a task that was designed to be completed based on time domain information alone.

The effect of speaker on accent identification performance did not correlate with speakers' foreign accent strength. As in Chapter 5, we therefore assume that speakers' non-native temporal patterns are speaker-individual to some extent. Chapter 7 provides an account of speaker-individual temporal features that vary between speakers in their native as well as non-native languages.

Speech Communication 86 (2017) 121–134

Contents lists available at ScienceDirect

Speech Communication

journal homepage: www.elsevier.com/locate/specom

# Listeners use temporal information to identify French- and English-accented speech

Marie-José Kolly[a,b,*], Philippe Boula de Mareüil[b], Adrian Leemann[c], Volker Dellwo[a]

[a] Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland
[b] Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur (LIMSI), CNRS, Université Paris-Saclay, Rue John von Neumann, 91405 Orsay Cedex, France
[c] Phonetics Laboratory, Department of Theoretical and Applied Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, United Kingdom

ARTICLE INFO

*Article history:*
Received 13 December 2015
Revised 12 November 2016
Accepted 22 November 2016
Available online 30 November 2016

*Keywords:*
Accent identification
Speech rhythm
Temporal cues
Segmental cues
Additivity of cues
Forensic phonetics

ABSTRACT

Which acoustic cues can be used by listeners to identify speakers' linguistic origins in foreign-accented speech? We investigated accent identification performance in signal-manipulated speech, where (a) Swiss German listeners heard native German speech to which we transplanted segment durations of French-accented German and English-accented German, and (b) Swiss German listeners heard 6-band noise-vocoded French-accented and English-accented German speech to which we transplanted native German segment durations. Therefore, the foreign accent cues in the stimuli consisted of only temporal information (in a) and only strongly degraded spectral information (in b). Findings suggest that listeners were able to identify the linguistic origin of French and English speakers in their foreign-accented German speech based on temporal features alone, as well as based on strongly degraded spectral features alone. When comparing these results to previous research, we found an additive trend of temporal and spectral cues: identification performance tended to be higher when both cues were present in the signal. Acoustic measures of temporal variability could not easily explain the perceptual results. However, listeners were drawn towards some of the native German segmental cues in condition (a), which biased responses towards 'French' when stimuli featured uvular /r/s and towards 'English' when they contained vocalized /r/s or lacked /r/.



## 1. Introduction

"Judging by your accent, you must be French" – people regularly engage in foreign accent identification tasks in everyday social interactions. Which acoustic cues are useful for such tasks? The question is particularly relevant when the origin of an individual has to be determined for legal cases, where forensic phoneticians or ear-witnesses establish a speaker's profile to reduce the number of potential suspects (Ellis, 1994; Köster, et al., 2012). Aside from forensic caseworkers, a number of governmental institutions conduct Linguistic Analyses for the Determination of the geographical Origin (LADO) of an individual. Here, an asylum seeker's claim to originate from a particular region is examined, when no valid identification documents are available (Baltisberger and Hubbuch, 2010). Foreign accent identification can be a crucial part of speaker profiling and LADO, as some individuals use second language speech to disguise their native language and thus their geographical origin (Cambier-Langeveld, 2010).

Foreign-accented speech contains a large number of specific features, and some of these are perceptually salient in terms of geographical origin. The most salient features indicative of a foreign accent are likely to be found on the segmental level (Boula de Mareüil, et al., 2004a; Boula de Mareüil, et al., 2008; Cunningham-Andersson and Engstrand, 1989; Flege and Port, 1981; Vieru, et al., 2011). /r/ in the Swiss German toponym *Zürich*, for example, is typically realized as a uvular trill [ʁ] or fricative [ʁ] by French speakers, and as an alveolar approximant [ɹ] by English speakers – as opposed to the Zurich Swiss German articulation of an alveolar trill [r] or tap [ɾ] (Werlen, 1980). Foreign-accented speech is characterized, to some extent, by interferences from the speakers' first language. Based on such interferences, for example in the /r/ realization, listeners can typically guess the native language (i.e., French, English, Swiss German) of the speaker.

In some adverse listening situations, access to segmental cues is reduced. One can think of speech that was recorded through a closed door, on a mobile telephone, or in a noisy environment, as

* Corresponding author at: Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, Plattenstrasse 54, 8032 Zurich, Switzerland.
*E-mail addresses:* marie-jose.kolly@uzh.ch (M.-J. Kolly), philippe.boula.de.mareuil@limsi.fr (P. Boula de Mareüil), al764@cam.ac.uk (A. Leemann), volker.dellwo@uzh.ch (V. Dellwo).

http://dx.doi.org/10.1016/j.specom.2016.11.006
0167-6393/

typically encountered in the domain of forensic phonetics: telephone speech is involved in 90% of forensic phonetic casework (Hirson, et al., 1995), and speech material for LADO, too, is often obtained over a landline network (Baltisberger and Hubbuch, 2010). Forensic caseworkers' decisions must most often rely on degraded segmental cues and/or on other cues. Here, speech prosodic information might play a crucial role: Listeners' ability to recognize words, for example, was shown to strongly deteriorate in noise, while their ability to recognize prosodic patterns remained unaffected by it (Van Zyl and Hanekom, 2011). However, adverse listening conditions often also reduce certain types of prosodic features, particularly features from the frequency domain. When speech is transmitted through a mobile telephone, for example, the frequency range is reduced to a frequency band between 350 and 3200 Hz (Künzel, 2001), measurements of vowel qualities are obscured (Byrne and Foulkes, 2004), and speakers' fundamental frequency tends to be higher due to speaking more loudly on the telephone (ibid.). *Temporal* cues are typically less affected by distortions of the speech signal as they occur in telephone speech (Chen, et al., 2005; et al., 2014). In the context of the present paper, we use the term *temporal* to refer to durations of speech segments, as this is the feature that we manipulated in our stimuli. Segment durations have an effect not only on segmental but also on suprasegmental timing patterns (van Santen and Shih, 2000).

Can listeners identify the origin of speakers based on temporal features of their non-native speech? A rationale for this idea comes from the domain of speech rhythm research – the study of the suprasegmental temporal organization of speech. Languages have been argued to differ in their rhythm (Abercrombie, 1967; Lloyd James, 1929; Pike, 1945). The acoustic features that allegedly correlate with the perception of speech rhythm remain to be fully determined, as rhythm metrics proposed in the literature were reported to be influenced not only by language (Dellwo, 2006; Grabe and Low, 2002; Ramus, et al., 1999) or dialect (Ferragne and Pellegrino, 2004; Leemann, et al., 2012; White and Mattys, 2007b), but also by factors such as speaker, sentence material, or annotator (Arvaniti, 2012; Dellwo, et al., 2015; Leemann et al., 2014; Vieru et al., 2011; Wiget et al., 2010). Numerous studies reported that listeners are sensitive to suprasegmental temporal information contained in speech (e.g. Pinet and Iverson, 2010; Quené and van Delft, 2010; Tajima, et al., 1997). Furthermore, listeners were reported to use such information to distinguish between languages (Nazzi, et al., 1998; Ramus and Mehler, 1999; Ramus, et al., 2003) or dialects (adults: White, et al., 2012; infants: White, et al., 2014). It is thus conceivable that suprasegmental temporal information might be a potential cue to foreign accents such as French-accented and English-accented German.

French and English differ in their suprasegmental temporal organization. For example, English features higher durational variability between prominent and less prominent syllables than French (Delattre, 1966; Fant et al., 1991). French and English also differ on the segmental temporal level: English, but not French, features distinctive vowel quantity and vowel reduction; English has more complex syllables and consonant clusters than French (Auer, 2001; Dauer, 1983; German shows similar temporal features as English in these examples). Speakers of both French and English produce longer vowels before voiced than before unvoiced consonants, but this effect is stronger for English speakers (Laeufer, 1992). These segmental temporal differences between the two languages may translate to differences in suprasegmental temporal structure as well (van Santen and Shih, 2000). For example, listeners were shown to perceive French as more regularly timed than English or German (Dellwo, 2008). Furthermore, some of the temporal patterns discussed are typically carried over to a non-native language (Arslan and Hansen, 1997; McAllister, et al., 2002). Voice Onset Time (VOT), for instance, is known to differ between French and English, and Hazan and Boulakia (1993) reported that bilingual speakers of French and English often produce VOT according to their dominant language. In conclusion, we start from the assumption that French-accented German and English-accented German differ in their segmental and suprasegmental temporal organization. We therefore hypothesize that listeners may be able to use such temporal features to identify the two accents.

The question whether particular foreign accents can be identified based on temporal cues has been studied only to a minor extent. Previous research on foreign accent identification more often than not featured material that contained a certain amount of frequency domain information in addition to temporal information: segment durations and intonation in prosody-transplanted speech (Boula de Mareüil and Vieru-Dimulescu, 2006); segment durations and degraded spectral features in 1-bit requantized speech (Kolly and Dellwo, 2014); temporal features of the amplitude envelope and degraded spectral features in 6-band noise-vocoded speech (Kolly and Dellwo, 2014); and temporal features of the amplitude envelope and of voicing in monotonized lowpass-filtered speech below 300 Hz (where some spectral features below 300 Hz may have been useful for accent identification; Kolly, et al., 2014). In this line of research, listeners were reported to respond at chance level when stimuli contained (almost) no spectral features, e.g. in 3-band noise-vocoded speech and in monotonized *sasasa*-speech (see below; Kolly and Dellwo, 2014). The signal conditions discussed preserve mainly temporal features and different degrees of rudimentary spectral information. Findings showed that accent identification performance decreased with higher degradation of spectral features. The outcome of this research can be interpreted in two ways: on the one hand, the additivity of cues may have played a role, where the combination of temporal and spectral features potentially boosted identification performance (Du et al., 2011; Hjalmarsson, 2011). Listeners might, for example, identify an accent because some rudimentary spectral information occurs at a specific (and expected) moment in time. If the temporal integrity of the signal were completely degraded, the same spectral information might be of less or no use to the listener. Similarly, if the spectral information were completely absent, the temporal information, still intact, may be of less or no use to a listener (Dellwo, 2010). On the other hand, temporal information alone might allow for foreign accent identification if it were presented in a signal condition that occurs in natural listening situations. In fact, 3-band noise-vocoded speech and *sasasa*-speech are highly distorted signals: The process of noise-vocoding replaces the source signal of speech with white noise (Shannon, et al., 1995), and, in the *sasasa*-experiment, every voiced interval was replaced with the same [a]-sound and every unvoiced interval with the same [s]-sound. 'Speech'-signals such as these do not occur in everyday listening situations. It thus seems plausible that, because of a lack of experience with such signals, listeners are not able to interpret the temporal information contained in them.

To test whether listeners rely on the additivity of temporal and spectral cues to identify foreign accents, we separated both cues contained in the 6-band noise-vocoded speech used by Kolly and Dellwo (2014). We conducted two perception experiments to investigate if listeners can identify foreign accents (a) based on temporal features alone (henceforth *timeOnly*), and (b) based on strongly degraded spectral features alone (henceforth *freqOnly*). To isolate temporal features for (a), and to eliminate temporal features for (b), we used a signal manipulation frequently referred to as 'prosody transplantation'. The method was introduced by Osberger and Lewitt (1979) and has mostly been applied to investigate the importance of temporal and/or fundamental frequency patterns for the intelligibility of deaf speakers (Maassen and Povel, 1985; Osberger and Lewitt, 1979) and the intelligibility and/or degree of accentedness in non-native speech (Holm, 2008;

Pinet and Iverson, 2010; Quené and van Delft, 2010; Rognoni and Busà, 2014; Tajima et al., 1997; Vitale, et al., 2014; Winters and O'Brien, 2013). Prosody-transplanted speech has also been used to investigate whether segmental or prosodic cues are more important to identify foreign accents; findings suggest that segmentals prevail in the identification of native vs. Arabic- or Kabyle-accented French (Boula de Mareüil et al., 2004a), whereas prosody plays more into the identification of Spanish-accented Italian vs. Italian-accented Spanish (Boula de Mareüil and Vieru-Dimulescu, 2006; Boula de Mareüil, et al., 2004b).

For the signal condition *timeOnly*, we transplanted segment durations of French- and English-accented German to native German, i.e., we modified German segment durations to match the segment durations of French- and English-accented German. This eliminated all spectral features of the foreign accents, while keeping the resulting stimuli fairly natural-sounding. For the signal condition *freqOnly*, we transplanted native German segment durations to French- and English-accented German, which eliminated all segmental and suprasegmental temporal information of the foreign accents. We then 6-band noise-vocoded the material in such a way that it contained the spectral information from 6-band noise-vocoded speech (Kolly and Dellwo, 2014). Apart from the fact that it allowed us to test effects of cue additivity, 6-band noise-vocoding was also performed to reduce spectral information, as it seemed plausible that intact spectral cues alone would lead to near-ceiling effects in perception experiments. A drawback of using the prosody transplantation and noise-vocoding approach is the artificiality of stimuli: the noise-vocoded speech of the *freqOnly* stimuli sounds highly unnatural; *timeOnly* speech sounds relatively natural but combines native frequency domain features with non-native temporal features, a hybrid signal that listeners also do not encounter in natural environments. However, this seems to be the ecologically most valid way of separating temporal and spectral features.

Our approach was (a) to test, in a perception experiment, whether listeners can recognize French- and English-accented German based on temporal features or spectral features of the foreign accents only, and (b) to investigate acoustic correlates that may explain listeners' behavior. In perception experiments, Swiss German listeners heard French- and English-accented *timeOnly* or *freqOnly* sentences and had to decide whether they heard a French or an English accent. We used a between-subjects design in which each signal condition was tested with different listeners, given that listeners may adapt to manipulated speech: Davis, et al. (2005), for example, reported that the intelligibility of noise-vocoded speech increased with training. In the context of the present study, a within-subjects design may have encouraged listeners to use their familiarization with the sentence, speaker and accent characteristics from, say, the *timeOnly* experiment when completing the task in the *freqOnly* experiment, resulting in artifacts, as such information would have been of no use to them. To allow for a comparison with previous experiments, we used the recordings and experiment design from Kolly and Dellwo (2014). A number of acoustic temporal measures were applied to unmanipulated speech and to our stimuli in order to verify that duration transplantation had the desired effect on the material. Furthermore, these acoustic temporal measures were used to explore potential acoustic correlates of listeners' accent identification performance.

## 2. Materials and methods

### 2.1. Subjects

A total of 40 native Swiss German listeners (16 male, 24 female) took part in the accent identification experiments. Listeners were University of Zurich students aged between 18 and 45 years (M = 23.30, SD = 4.37). None of them reported hearing disorders or problems with sight. Due to listeners' age, origin and educational level, we assumed a comparable level of familiarity with French and English speakers of German. Likewise, we presupposed similar levels of proficiency in French and English, as French is usually introduced as a second and English as a third language in Swiss German schools: Subjects had studied French and English for about 11 and 6 years, respectively. Before starting university studies, Swiss German students such as our subjects pass an exam called *Maturität* (*Baccalaureate*), for which their proficiency in French and English is expected to correspond to B2–C1 according to the Common European Framework for Languages (Council of Europe, 2013; Erziehungsdirektion des Kantons Bern, 2009). At university, students tend to use English more than French.

For our between-subject design, listeners were randomly attributed to two groups. We tested 20 listeners (10 male, 10 female) with the signal condition *timeOnly* and 20 listeners (6 male, 14 female) with the signal condition *freqOnly*.

### 2.2. Materials

#### 2.2.1. Speakers

We collected Standard German speech from 18 speakers: three male and three female speakers for each language (French, English and Zurich German). Speakers' age ranged between 23 and 56 years (M = 30.78, SD = 8.02). The Zurich German speakers grew up in the city of Zurich; the French speakers in the French-speaking part of Switzerland; the English speakers in the US or in Canada, one female speaker in the UK (their English varieties feature similar durational patterns, for instance vowel reduction; Grenon and White, 2008; Shearme and Holmes, 1961; Tiffany, 1959).

Native Standard German speech for duration transplantation was obtained from Zurich German speakers, as our listeners were mostly Zurich German, too. In diglossic German-speaking Switzerland, dialects are used mainly for verbal communication, whereas Standard German is mainly used in the written form and in more formal oral situations (Ferguson, 1959; Kolde, 1981). The pronunciation of /r/ in Swiss Standard German is variable (Hove, 2002): some speakers produce an alveolar [r] or [ɾ], the variant present in most of the Swiss German dialects (Werlen, 1980); others produce /r/ as a uvular trill [ʀ] or fricative [ʁ]. In specific phonotactic positions, certain speakers may vocalize /r/ to schwa [ɐ], particularly in post-vocalic contexts, which corresponds to the Standard German system (Kohler, 1990). The Zurich German speakers recorded for the present experiments all used uvular as well as vocalized /r/ variants in their Standard German.

The Zurich German speakers used Standard German on a regular basis. French and English speakers self-assessed their proficiency in German using the Common European Framework for Languages (Council of Europe, 2013). French speakers' proficiency ranged between B1 and B2, English speakers' between A1 and B2. The origin and strength of their foreign accent was rated by 16 listeners (9 male, 7 female) in natural speech, on a 5-point scale (1 = very strong accent, 2 = strong accent, 3 = medium accent, 4 = slight accent, 5 = no accent). Listeners' age ranged between 20 and 36 years (M = 26.25, SD = 5.20). None of the listeners was part of the group of subjects presented in Section 2.1. We constructed a linear model of *accent strength* as a function of *accent* and found no significant differences in *accent strength* between the French and the English speaker group (LM: $F(1,10) = 0.39$, $p = 0.55$; French speakers: M = 2.86, SD = 0.19; English speakers: M = 2.67, SD = 0.71; cf. Section 2.6 for details on statistical analyses). We further found their foreign accents to be recognized with high performance, in natural speech, as measured by *A'* (M = 0.95, SD = 0.03). This illustrates that speakers provided typical examples of French- and

**Table 1**
Ranking of male and female speakers according to articulation rate as measured by *ratePeak*.

| Gender | French speakers | Zurich German speakers | English speakers |
|---|---|---|---|
| Male | FR04 ($M = 5.41$, SD = 0.85) | ZH07 ($M = 5.55$, SD = 0.42) | EN01 ($M = 5.50$, SD = 0.43) |
| | FR01 ($M = 4.82$, SD = 0.71) | ZH14 ($M = 5.21$, SD = 0.54) | EN06 ($M = 5.26$, SD = 0.81) |
| | FR10 ($M = 4.14$, SD = 0.30) | ZH15 ($M = 5.11$, SD = 0.47) | EN07 ($M = 5.14$, SD = 0.54) |
| Female | FR05 ($M = 5.09$, SD = 0.35) | ZH69 ($M = 5.63$, SD = 0.39) | EN03 ($M = 4.98$, SD = 0.45) |
| | FR03 ($M = 5.08$, SD = 0.58) | ZH71 ($M = 5.37$, SD = 0.52) | EN02 ($M = 4.83$, SD = 0.59) |
| | FR08 ($M = 4.75$, SD = 0.43) | ZH70 ($M = 5.22$, SD = 0.58) | EN04 ($M = 4.17$, SD = 0.63) |

English-accented speech, and corresponds to the judgement of expert phoneticians (authors).

#### 2.2.2. Reading materials and recordings

All speakers read a list of 18 Standard German sentences, which varied between 12 and 16 syllables (cf. Appendix). Prior to the recording, speakers familiarized themselves with the materials by reading the sentences aloud. The French and English speakers were recorded in a quiet room using a *Fostex FR-2LE* solid-state recorder (48 kHz; 16 bit) and a *Sennheiser MKE 2p-c* clip-on microphone. The Zurich German speakers were recorded in a sound-treated booth using a *Neumann STH-100* transducer microphone (44.1 kHz; 16 bit). We selected a different set of 9 sentences from each French and English speaker to avoid identical sentence sets for all speakers and thus to obtain more variability of linguistic material in the experiment (Kolly and Dellwo, 2014). The experiment contained 108 sentences in total (2 accents × 6 speakers × 9 sentences): Each of the 18 Standard German sentences appeared six times in the experiment, three times read by a French speaker and three times read by an English speaker.

#### 2.2.3. Segmentation

The 108 non-native sentences and their native German counterparts were segmented, on a phonetic level, by a trained phonetician (first author), using Praat (Boersma and Weenink, 2014). Segmentation and labeling decisions were based on visual inspection of waveforms and spectrograms, and on auditory criteria. All interval boundaries were placed at positive zero-crossings. In order to obtain an optimal transplantation of durational patterns, diphthongs and affricates were segmented into their components, glottal stops or laryngealized parts were treated as individual segments, and silent pauses were annotated without the application of a particular duration threshold. However, stops were not divided in separate closure and release sections.

### 2.3. Stimuli

We chose to transplant segment durations rather than syllable durations (e.g. Maassen and Povel, 1985; Osberger and Lewitt, 1979; Winters and O'Brien, 2013) since French- and English-accented German may differ on a very detailed durational level (cf. Section 1). Furthermore, segmental durations have been suggested to be an important cue for foreign accent identification (Kolly and Dellwo, 2014). Segment durations of the speech material read by each particular French and English speaker were therefore transplanted to material read by a native speaker (*timeOnly*) and vice versa (*freqOnly*).

Since we transplanted durational features, speaker pairs (French-Zurich German; English-Zurich German) were built according to a gender-specific ranking of articulation rate (cf. Table 1), as measured by *ratePeak* (cf. Section 2.4). In doing so, we avoided an extreme stretching of segments – which may result in artifacts such as chirp or whistle sounds – wherever possible (Quené and van Delft, 2010). For the signal condition *timeOnly*, for example, segment durations of FR04 (and those of EN01) were transplanted to ZH07.

After the segmentation (cf. Section 2.2.3) we checked whether the matching versions of each sentence from each speaker pair (e.g. speakers FR04 and ZH07, sentence 03) were segmented into the same number of intervals, a prerequisite for the transplantation of segment durations. The number of intervals differed between the versions if either the number of segments, or the number of silent pauses was different. If only one version of a sentence featured a silent pause at a specific position, we introduced a silent part of the same length and at the same position to the other version, and added an interval to its segmentation. The silent part that was introduced was taken from the (silent) start or end of the sentence into which it was introduced, in order to obtain a maximally natural auditory effect (Pettorino and Vitale, 2012). In cases where the segment count was different, we merged intervals in the version that contained a higher number of segments, which resulted in some intervals containing multiple segments (cf. Fig. 1). Intervals were merged according to syllable or phoneme boundaries, such that durational features of a syllable or phoneme would be transplanted to the same syllable or phoneme of the matching version of the sentence (Tajima et al., 1997). Typical examples for situations where intervals were merged are the following:

- Elisions:
  - Some native German speakers elided the schwa before a sonorant in unstressed syllables (e.g. *Regen* [ˈʁeːgn] vs. [ˈʁeːgən] 'rain'). In such cases, the schwa and the following sonorant of the French or English speaker's sentence were merged into a single interval, as exemplified in Fig. 1.
  - Some French or English speakers elided linking elements between the two components of a German compound (e.g. *Zahlungsbilanz* [ˈtsaːlʊŋbiˌlants] vs. [ˈtsaːlʊŋsbiˌlants] 'balance of payment'). In such cases, the linking element and the preceding phone of the Zurich German speaker's sentence were merged into a single interval.
- Epentheses: Some French and English speakers produced a velar plosive after velar nasals (e.g. *lange* [ˈlaŋgə] vs. [ˈlaŋə] 'long'). In such cases, the nasal and the following plosive were merged into a single interval.

Fig. 2 illustrates the signal processing steps undertaken to obtain stimuli for *timeOnly*: Segment durations of the native version of a sentence were modified with segment durations of its non-native counterpart. Native German speech intervals were therefore either stretched or compressed by means of Pitch Synchronous Overlap and Add (PSOLA) resynthesis, using a Praat script adapted from Boula de Mareüil and Vieru-Dimulescu (2006). The speech signal, albeit carrying some artifacts due to the stretching of particular segments, is still intelligible and rather natural. To obtain stimuli for *freqOnly*, segment durations of the non-native version of a sentence were modified with segment durations of its native counterpart. Sentences were subsequently 6-band noise-vocoded. We divided the speech signal into six logarithmically-spaced frequency bands. We used the same respective cutoff frequencies to filter white noise. The amplitude envelope was extracted from each

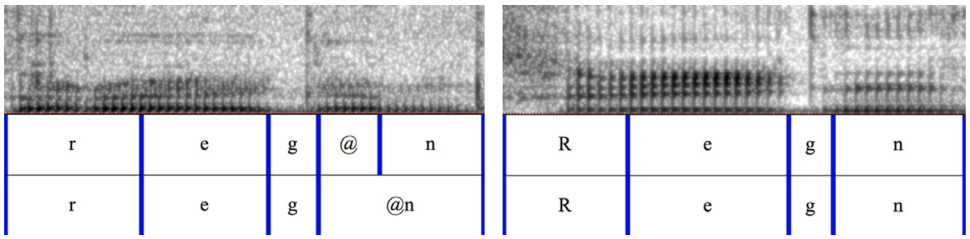



**Fig. 1.** Annotation of segments (tier 1) and parallel annotation of two matching versions of a sentence resulting in merged segments (tier 2) for an English-accented (left spectrogram and annotation) and a native German (right spectrogram and annotation) token of *Regen* 'rain'.

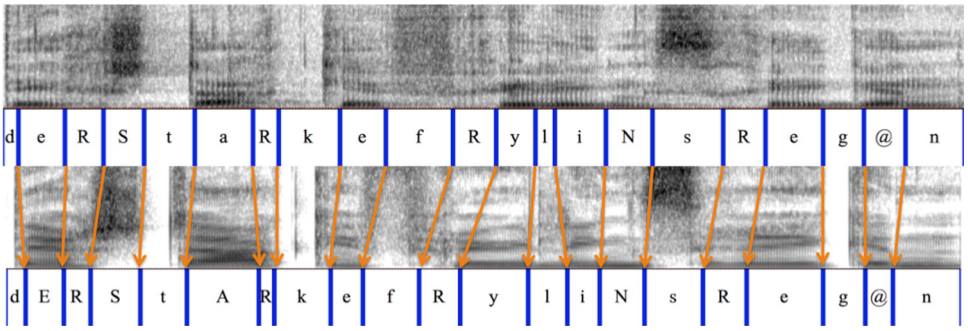



**Fig. 2.** Modification of native German segment durations (bottom spectrogram and annotation) with French segment durations (top spectrogram and annotation). The phrase reads *der starke Frühlingsregen* 'the strong spring rain'.

speech band and multiplied with the corresponding noise band. The six noise bands were summed up to obtain 6-band noise-vocoded speech (cf. Kolly and Dellwo, 2014, for more detail). All stimuli were scaled to an intensity of 70 dB.

### 2.4. Temporal measures applied

In the following, we present a number of acoustic measures that were applied to the material (i) to describe our stimuli and therefore verify what effect the duration transplantation may have had on certain durational characteristics of the material (cf. Section 3.2.1) and (ii) to explore potential acoustic correlates of listeners' identification performance (cf. Section 3.2.2). For this, we applied five different types of temporal measures to the natural and the duration transplanted speech: (1) measures of articulation rate, (2) pausing measures, and (3) a number of rhythm metrics based on the durational variability (3a) of vocalic and consonantal intervals, (3b) of voiced and voiceless intervals and (3c) of intervals between peaks in the amplitude envelope.

(1) Measures of articulation rate:

- *rateCV*, the number of consonantal and vocalic intervals per second (Dellwo, 2008);
- *ratePeak*, the number of automatically detected peaks in the amplitude envelope (Dellwo et al., 2012; Mermelstein, 1975), which roughly corresponds to the number of syllables, per second.

(2) Measures of pausing (Bosker et al., 2014; Cucchiarini et al., 2002; de Jong et al., 2013; Künzel, 1997):

- *pauseNbr*, the number of silent pauses;
- *pauseDur*, silent pause durations.

(3a) Rhythm metrics based on durational features of vocalic and consonantal intervals (derived from segmentation):

- *%V*, the percentage of time over which speech is vocalic (Ramus et al., 1999);
- *varcoVln*, the rate-normalized standard deviation of vocalic interval durations (*varcoV*: White and Mattys, 2007a), calculated on log-transformed interval durations;
- *nPVI_V*, the rate-normalized average difference between consecutive vocalic interval durations (Grabe and Low, 2002);
- *varcoC*, the rate-normalized standard deviation of consonantal interval durations (Dellwo, 2006);
- *nPVI_C*, the rate-normalized average difference between consecutive consonantal interval durations (Grabe and Low, 2002).

(3b) Rhythm metrics based on durational features of voiced and unvoiced intervals (automatically calculated using the default pitch detection algorithm in Praat):

- *%VO*, the percentage of time over which speech is voiced (Dellwo, et al., 2007);
- *varcoVOln*, the rate-normalized standard deviation of voiced interval durations (*varcoVO*: Dellwo et al., 2007), calculated on log-transformed interval durations;
- *nPVI_VO*, the rate-normalized average difference between consecutive voiced interval durations (Dellwo et al., 2007);
- *varcoUV*, the rate-normalized standard deviation of unvoiced interval durations (Dellwo et al., 2007);
- *nPVI_UV*, the rate-normalized average difference between consecutive unvoiced interval durations (Dellwo et al., 2007).

(3c) Rhythm metrics based on durational features of intervals between automatically detected peaks in the amplitude envelope (one peak per vocalic segment):

- *varcoPeak*, the rate-normalized standard deviation of interval durations between automatically extracted amplitude peaks (Dellwo et al., 2012);
- *nPVI_Peak*, the rate-normalized average difference between consecutive interval durations between automatically extracted amplitude peaks (Dellwo et al., 2012).

The measures *varcoV* (White and Mattys, 2007a) and *varcoVO* (Dellwo et al., 2007) were calculated based on log-transformed interval durations, since the distributions of vocalic and voiced intervals were strongly positively skewed. Temporal measures were calculated sentence-by-sentence using the Praat plugin *Duration Analyzer* (available at http://www.pholab.uzh.ch/static/volker/software/plugin_durationAnalyzer.zip).

### 2.5. Procedure

Listeners were tested in a quiet room at the University of Zurich using a laptop computer. They heard stimuli over high-quality closed *Beyerdynamics DT 770 PRO* headphones, and stimulus order was randomized for each listener. Listeners tested for *freqOnly* heard strongly distorted speech; they were thus presented with sentence transcripts corresponding to each acoustic stimulus, which allowed them to parse the acoustic information (Davis et al., 2005). Sentence transcripts were presented on the computer screen two seconds prior to the acoustic stimulus, and remained on the screen during stimulus presentation. Listeners tested in the *timeOnly* signal condition were not given sentence transcripts, as the stimuli presented were readily intelligible. We cannot exclude that the display of sentence transcripts distracted listeners' attention from the acoustic signal in *freqOnly*; listeners tested with *timeOnly*, on the other hand, could focus their entire attention on the acoustic stimulus. Prior findings by Kolly and Dellwo (2014) suggest, however, that this potentially distracting effect is small compared to the gain from listeners being aware of the sentence content: Listeners identified accents above chance in 6-band noise-vocoded speech when the acoustic stimuli were presented with sentence transcripts, but not when they were missing.

Listeners were instructed as follows: they would hear Standard German sentences spoken by French and English speakers and they would have to decide, for each sentence, whether they heard French- or English-accented German, and how confident they were concerning their response. They were encouraged to respond intuitively. Listeners tested in the *freqOnly* signal condition were additionally informed that they would hear manipulated speech and that they would be able to read the sentence corresponding to the acoustic stimulus on the computer screen. They responded using a binary forced choice experiment interface presented over the Praat demo window function (comparable Praat plugin available at http://www.pholab.uzh.ch/static/volker/software/plugin_BFC_Experiment.zip). After each stimulus presentation, a response window appeared with the question *Französischer oder englischer Akzent?* 'French or English accent?'. Below this text, there were two large grey rectangles titled *Französisch* and *Englisch*. Each of them contained three small blue rectangles that read *sicher* 'confident', *weiss nicht recht* 'not confident', and *nur geraten* 'only guessing'. Listeners clicked on one of the blue rectangles, indicating whether they judged the stimulus as being French- or English-accented German. At the same time, they indicated their confidence level for each stimulus on a 3-point scale. Before the beginning of the experiment, listeners were familiarized with the experiment interface and with manipulated speech through the display of two randomly selected stimuli. The experiment, including instructions, lasted about 20 min and listeners were paid 10 Swiss Francs for their participation.

### 2.6. Data analysis and statistical analyses

Based on listeners' responses, we computed a measure of sensitivity derived from Signal Detection Theory (Green and Swets, 1966) in order to capture listeners' accent identification performance while cancelling out response bias. The non-parametric sensitivity measure *A'* and the corresponding measure of response bias, $B''_D$, were calculated following Donaldson (1992). We arbitrarily attributed French-accented German to be signal and English-accented German to be noise; responding 'French' to a French-accented stimulus was thus defined to be a *hit*, whereas responding 'English' to an English-accented stimulus was a *correct rejection*. The two error types, *false alarm* and *miss* were thus the response 'French' to an English-accented stimulus and the response 'English' to a French-accented stimulus, respectively. *A'* ranges from 0 to 1, with chance level at 0.5: a listener with an *A'*-value of 0 shows systematic confusion of the stimuli, i.e., responded incorrectly to all stimuli; an *A'*-value of 1 indicates perfect sensitivity. The values for bias ($B''_D$) range from −1 to 1, 0 indicating no bias, negative values indicating bias towards the response 'French' and positive values indicating bias towards 'English'. An alternative to *A'* and $B'_D$ are the measures *d'* and $\beta$ respectively, which assume underlying normal distributions of hit and false alarm rates. As we obtained comparable results with *d'* and, with one exception (cf. Section 3.2.3), for $\beta$, we do not report these values. When presenting effects of *accent* and *speaker*, it was not possible to report *A'* as we were interested in the responses to each of the two signal types separately. This is why we reported the percentage of correct responses, *%correct*, instead.

Statistical analyses were performed using R software (R Core Team, 2013). To test the magnitude of listeners' sensitivity, we calculated two-sided one-sample t-tests. To test for the effect of different factors on listeners' sensitivity, we constructed linear models (LM). Wherever possible, we calculated linear mixed effects models with *speaker gender, accent* and *signal condition* as fixed effects and *speaker, sentence* and *listener* as random intercepts (LME; R-package: *lme4*; Bates and Maechler, 2009). We also used linear mixed effect models for acoustic analyses of speech production. Here, our models included *gender, accent* and *transplantation* as fixed effects, *speaker* and *sentence* as random intercepts. Effects were tested by comparing a full model, which included the factor in question, to a reduced model, in which the factor was not included. Model comparison was performed using standard likelihood ratio tests (R-code: anova(full_model, reduced_model). We report AIC (Akaike Information Criterion) values for the relative goodness of fit of LMEs (Kliegl, et al., 2011). For multiple comparisons, we applied the Tukey method, using the R-package *multcomp*. For correlations, we report Spearman's correlation coefficient. We assumed an $\alpha$-level of 0.05.

## 3. Results

We present results on listeners' accent identification performance in *timeOnly* and *freqOnly* signal conditions in Section 3.1.1, and Section 3.1.2 compares these results with findings on accent identification performance when both types of cues are combined, in *time+freq* (adapted from Kolly and Dellwo, 2014). In Section 3.2, we investigate potential acoustic correlates of the perceptual results: To verify that our stimuli convey temporal or spectral information of the foreign accents only, we describe the acoustic features of the stimuli in Section 3.2.1. Section 3.2.2 investigates whether acoustic temporal features of the *timeOnly* stimuli may explain listeners' identification performance. In Section 3.2.3, we explore how the native German segmental content may have biased listeners' responses in the *timeOnly* condition.

### 3.1. Results from the perception experiments

#### 3.1.1. Temporal cues and spectral cues in foreign accent identification

To test the magnitude of listeners' sensitivity, we calculated *A'* for each listener ($n=40$). A boxplot of *A'* for each *signal condition* is presented in Fig. 3 (left graph). One-sample t-tests showed that sensitivity was significantly above chance for *timeOnly* ($t(19)=2.42$, $p<0.05^*$) as well as for *freqOnly* ($t(19)=7.69$, $p<0.001^*$). We found a significant effect of *condition*: listeners identified accents with greater performance in *freqOnly* ($M=0.63$, $SD=0.08$) than in *timeOnly* ($M=0.54$, $SD=0.07$; LM: $F(1,38)=15.85$, $p<0.001^*$).

Fig. 3 (right graph) shows one boxplot of $B''_D$ per *signal condition*, indicating listeners' response bias. One-sample *t*-tests showed that listeners were significantly biased towards the response 'French' for *freqOnly* ($t(19)=-2.40$, $p<0.05^*$), but not for

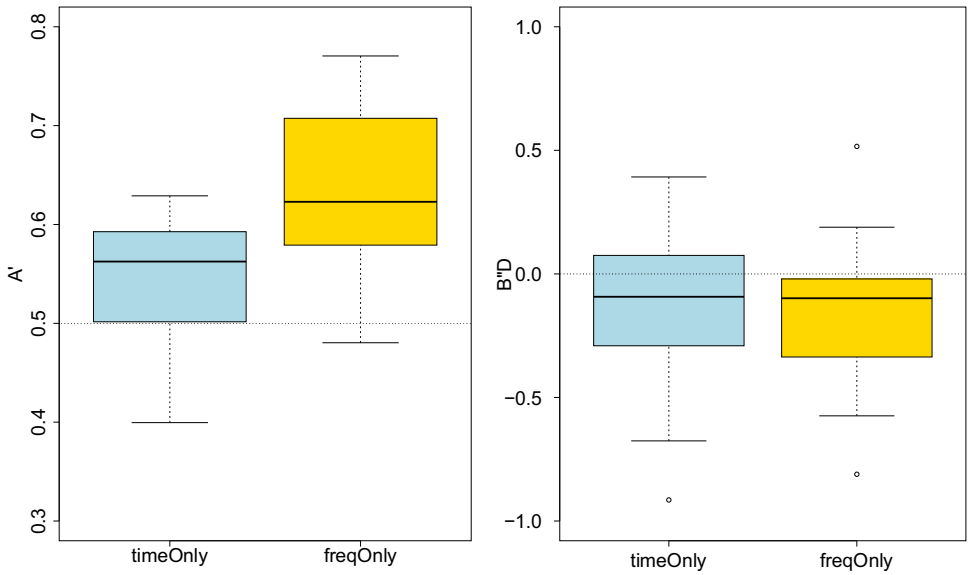



**Fig. 3.** Boxplots of listeners' accent identification performance as measured by *A'* (left graph) and listeners' response bias as measured by $B''_D$ (right graph), by *signal condition*. The dotted lines indicate performance at chance level and no bias, respectively.

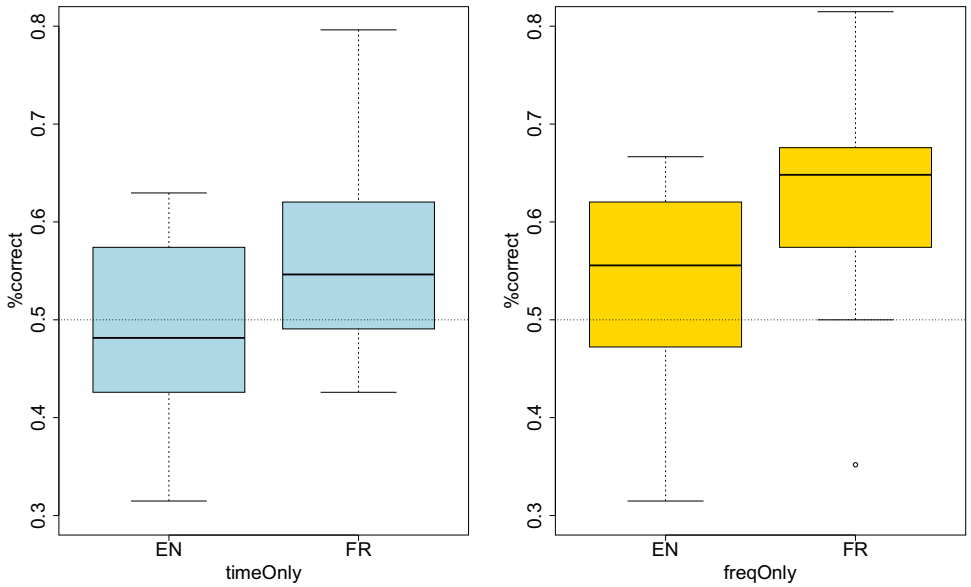



**Fig. 4.** Boxplots of listeners' accent identification performance as measured by *%correct*, per *accent*, for the signal conditions *timeOnly* (left graph) and *freqOnly* (right graph). The dotted lines indicate performance at chance level.

*timeOnly* (t(19)=−2.02, $p=0.06$). Listeners' bias did not differ significantly between *timeOnly* ($M=-0.15$, SD=0.33) and *freqOnly* ($M=-0.16$, SD=0.30; LM: F(1,38)= 0.01, $p=0.93$).

To test for the effect of *accent*, we calculated *%correct* for each listener's response to each accent ($n=80$: 2 accents × 40 listeners; as we investigated *accent* effects for each signal condition separately, we performed a Bonferroni-adjustment: 0.05/2=0.025). Boxplots of *%correct* by *accent* and *signal condition* are shown in Fig. 4. French accents were identified with significantly higher performance than English accents in *timeOnly* (LM: F(1,38)= 7.00, $p<0.025^*$; French: $M=0.57$, SD=0.10, English: $M=0.48$, SD=0.11) as well as in *freqOnly* (LM: F(1,38)=7.33, $p<0.025^*$; French: $M=0.62$, SD=0.10; English: $M=0.54$, SD=0.10).

To test for the effect of *speaker*, we calculated *%correct* for each listener's response to each speaker's sentences ($n=480$: 12 speakers × 40 listeners) and constructed an LME of *%correct* with *speaker gender, signal condition* and *accent* as fixed effects, a by-*speaker* random slope on *signal condition*, and random intercepts of *speaker* and *listener*. We obtained a significant effect of speaker ($\chi^2(3)=106.91$, AIC=−138.95, $p<0.001^*$). There was no correlation between speakers' strength of foreign accent (cf. Section 2.2.1) and the identification of their accent in either condition (*timeOnly*, $r=-0.21$, p=0.66; *freqOnly*, $r=0.23$, $p=0.33$). To test for the *sentence* effect, we calculated *A'* for each listener's response to each sentence ($n=720$: 18 sentences × 40 listeners) and constructed an LME of *A'* with *signal condition* as fixed effect and random intercepts on *sentence* and *listener*. There was no effect of sentence. Furthermore, listeners' confidence was found not to be significantly affected by *signal condition* (LM: F(1,38)=0.23, $p=0.64$) or *accent* (LM: F(1,78)=0.83, $p=0.37$).

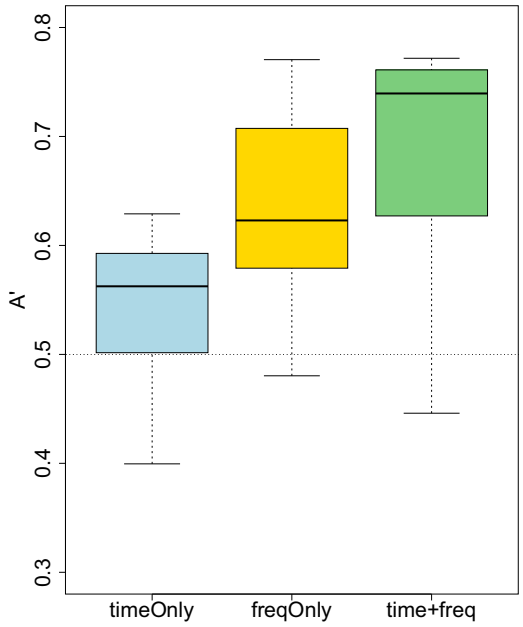



**Fig. 5.** Boxplots of listeners' accent identification performance as measured by *A'*, by *signal condition*, including data from the condition *time+freq* (adapted from Kolly and Dellwo, 2014). The dotted line indicates performance at chance level.

### 3.1.2. Additivity of temporal and spectral cues in foreign accent identification

Fig. 5 shows boxplots of *A'* for *timeOnly* (light blue) and *freqOnly* (yellow) in comparison to *time+freq* (green) adapted from Kolly and Dellwo (2014). *Time+freq* contained 6-band noise-vocoded speech with the original, non-native durations, thus featuring both the cues from *timeOnly* and *freqOnly* combined. Results showed a significant overall effect of *condition* (LM: $F(2,48)=9.96$, $p<0.001^*$). Post-hoc multiple comparisons revealed that *timeOnly* was significantly different from *freqOnly* ($p<0.01^*$) and *time+freq* ($p<0.001^*$); *freqOnly* and *time+freq* did not differ from each other significantly, however ($p=0.45$). Descriptively, *time+freq* yielded the highest *A'*-values ($M=0.67$, $SD=0.13$), followed by *freqOnly* ($M=0.63$, $SD=0.08$) and *timeOnly* ($M=0.54$, $SD=0.07$).

## 3.2. Results from the acoustic analyses

### 3.2.1. Acoustic measures of temporal variability in the stimuli

To test whether our material contains the intended acoustical information, we explored which temporal information is contained in the *timeOnly* stimuli, and tested whether *freqOnly* stimuli do in fact contain spectral information alone. To do this, we compared temporal patterns of French- and English-accented German in natural speech and duration transplanted speech. We hereby only applied rhythm metrics of the type (3b), voicing measures, and of the type (3c), peak measures (cf. Section 2.4): when transplanting segment durations, we automatically also copy temporal patterns such as articulation rate, pausing, as well as vocalic and consonantal interval durations. Measures of the type (1)–(3a) are thus not subject to change after duration transplantation. However, when transplanting segment durations to obtain *timeOnly* stimuli, voicing temporal patterns and the location of peaks in the amplitude envelope may only be captured to some extent: The proportion of voicing in individual segments and the location of amplitude peaks are known to differ between languages (voicing: Dellwo et al., 2007; amplitude peaks: Tilsen and Arvaniti, 2013) and speakers (voicing and amplitude peaks: Dellwo et al., 2015; Leemann et al., 2014). We therefore expect these features to be affected by duration transplantation to some extent. Furthermore, non-native speech is often characterized by L1-interference in voicedness, which is why voicing temporal patterns may be a useful cue in the perception task, if French- and English-accented *timeOnly* stimuli were to differ in this feature (Flege and Port, 1981; Hazan and Boulakia, 1993; Leemann, 2011; Neuhauser, 2011; Schmid, 2012; Vieru et al., 2011). In the *freqOnly* stimuli, voicing cues were absent due to 6-band noise-vocoding. However, it is important to examine that the French- and English-accented *freqOnly* stimuli do not differ in amplitude peak durational patterns, as these stimuli are intended to carry spectral cues only.

*3.2.1.1. Temporal patterns in timeOnly stimuli.* Results in Table 2 reveal that four out of five of the applied voicing measures were significantly affected by duration transplantation. Only *varcoVOln* did not differ before and after duration transplantation. The variability of intervals between amplitude peaks, however, seemed to be unaffected by duration transplantation.

In the case of *%VO*, we also observed a (marginally) significant effect of *accent* and, for *%VO* as well as *nPVI_VO*, a significant interaction of *transplantation* and *accent*. Simple effects for *%VO* ($\chi^2(1)=10.09$, $p<0.01^*$, AIC = 733.7; Bonferroni-adjustment: $0.05/2=0.025$) as well as for *nPVI_VO* ($\chi^2(1)=11.61$, $p<0.001^*$, AIC = 935.2) showed an effect of *transplantation* in French-accented speech only. Simple effects of *accent* revealed no significant difference between French- and English-accented German in natural or in transplanted speech for neither metric. Fig. 6 illustrates a descriptive (but non-significant) difference between voicing temporal patterns of the two accents in natural speech (FR vs. EN, natural), which vanishes in transplanted speech (FR vs. EN, *timeOnly*): *%VO* was higher in French ($M=73.95$, $SD=9.17$) than in English ($M=69.67$, $SD=6.91$) natural speech; *nPVI_VO* was lower in French ($M=66.09$, $SD=17.67$) than in English ($M=72.78$, $SD=17.39$) natural speech. Fig. 6 further illustrates, for a selection of the durational measures presented in Table 2, that most voicing measures were affected by duration transplantation, whereas the peak measures were not. For example, natural French-accented German exhibits significantly lower values for *nPVI_VO* than duration transplanted French-accented German (natural vs. *timeOnly*, FR). However, there is no such difference regarding the measure *nPVI_Peak*. Based on these results, we conclude that listeners could make little or no use of voicing temporal cues or amplitude peak temporal cues for identifying accents in the signal condition *timeOnly*.

*3.2.1.2. Temporal patterns in freqOnly stimuli.* As explained in Section 3.2.1, voicing cues are absent from the *fre-qOnly* stimuli due to noise vocoding. Therefore, only amplitude peak durational measures were applied to these stimuli in order to verify that they contain only frequency domain cues of the foreign accents.

Table 3 shows that neither of the applied peak durational measures in native German speech was significantly affected by duration transplantation. Therefore, *freqOnly* stimuli carry spectral information of the non-native accents, and temporal information of native German, as intended. Furthermore, we found no effect of *accent*, i.e., no difference between French-accented German with native German segment durations and English-accented German with native segment durations. Listeners thus had no durational cues available to complete the perceptual task in the signal condition *freqOnly*.

### 3.2.2. The influence of acoustic measures of temporal variability in foreign accent identification

We calculated correlations of listeners' accent identification performance – as measured by *%correct* – and 16 acoustic measures of temporal variability: two measures of articulation rate (measures of type (1), cf. Section 2.4), two pausing measures (type (2)), five measures of the durational variability of vocalic and consonantal

**Table 2**
Summary of the statistics for the tested voicing and peak measures in non-native natural speech and *timeOnly* stimuli. Acoustic measures are ordered according to the magnitude of the effect of *transplantation*.

| Temporal measure | Factor | Result | | |
|---|---|---|---|---|
| *nPVI_VO* | *transplantation* | $\chi^2(2)=11.66$ | $p<0.01^*$ | AIC=1872.1 |
| (voicing measure) | *accent* | $\chi^2(2)=4.42$ | $p=0.11$ | AIC=1872.1 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=3.84$ | $p=0.05^*$ | AIC=1872.1 |
| *%VO* | *transplantation* | $\chi^2(2)=11.38$ | $p<0.01^*$ | AIC=1423.7 |
| (voicing measure) | *accent* | $\chi^2(2)=5.97$ | $p=0.05^*$ | AIC=1423.7 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=4.25$ | $p<0.05^*$ | AIC=1423.7 |
| *varcoUV* | *transplantation* | $\chi^2(2)=10.30$ | $p<0.01^*$ | AIC=−114.88 |
| (voicing measure) | *accent* | $\chi^2(2)=1.47$ | $p=0.48$ | AIC=−114.88 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=1.46$ | $p=0.23$ | AIC=−114.88 |
| *nPVI_UV* | *transplantation* | $\chi^2(2)=9.27$ | $p<0.01^*$ | AIC=1955.5 |
| (voicing measure) | *accent* | $\chi^2(2)=0.87$ | $p=0.65$ | AIC=1955.5 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=0.74$ | $p=0.39$ | AIC=1955.5 |
| *varcoPeak* | *transplantation* | $\chi^2(2)=1.44$ | $p=0.49$ | AIC=−435.61 |
| (peak measure) | *accent* | $\chi^2(2)=0.97$ | $p=0.62$ | AIC=−435.61 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=0.97$ | $p=0.33$ | AIC=−435.61 |
| *nPVI_Peak* | *transplantation* | $\chi^2(2)=1.08$ | $p=0.58$ | AIC=1739.4 |
| (peak measure) | *accent* | $\chi^2(2)=3.39$ | $p=0.18$ | AIC=1739.4 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=0.95$ | $p=0.33$ | AIC=1739.4 |
| *varcoVOln* | *transplantation* | $\chi^2(2)=0.62$ | $p=0.73$ | AIC=−269.11 |
| (voicing measure) | *accent* | $\chi^2(2)=0.23$ | $p=0.89$ | AIC=−269.11 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=0.08$ | $p=0.78$ | AIC=−269.11 |

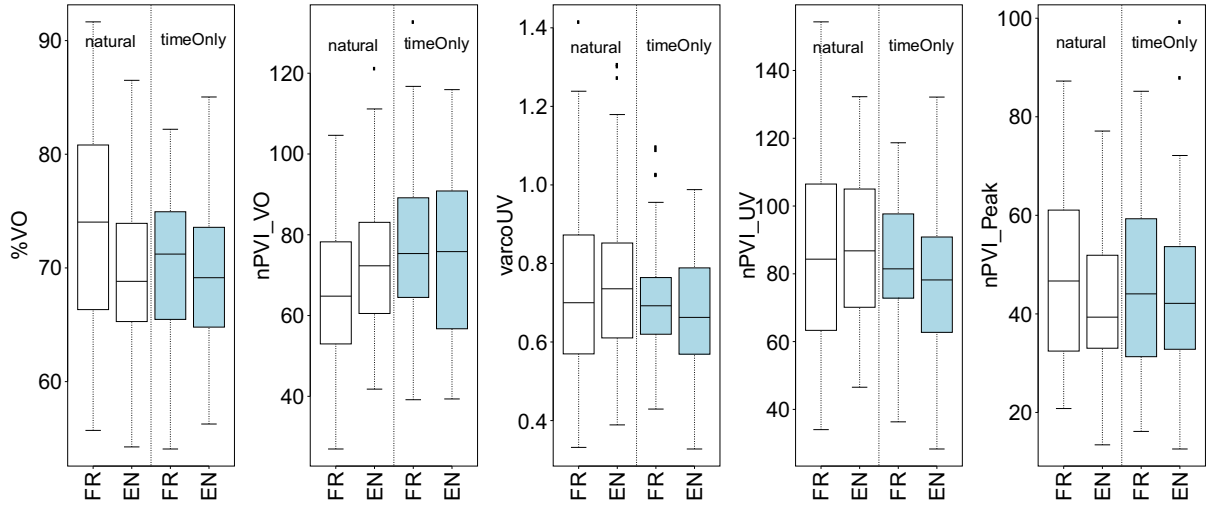



**Fig. 6.** Boxplots of *%VO, nPVI_VO, varcoUV, nPVI_UV* and *nPVI_Peak* in non-native natural speech (white) and in the *timeOnly* stimuli (light blue), for French-accented and English-accented speech. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article).

**Table 3**
Summary of the statistics for the tested peak measures in native natural speech and *freqOnly* stimuli. Acoustic measures are ordered according to the magnitude of the effect of *transplantation*.

| Temporal measure | Factor | Result | | |
|---|---|---|---|---|
| *varcoPeak* | *transplantation* | $\chi^2(2)=3.56$ | $p=0.17$ | AIC=−529.36 |
| (peak measure) | *accent* | $\chi^2(2)=0.19$ | $p=0.91$ | AIC= −529.36 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=0.17$ | $p=0.69$ | AIC=−529.36 |
| *nPVI_Peak* | *transplantation* | $\chi^2(2)=2.17$ | $p=0.34$ | AIC=1681.9 |
| (peak measure) | *accent* | $\chi^2(2)=0.09$ | $p=0.96$ | AIC=1681.9 |
| | *accent* ∗ *transplantation* | $\chi^2(1)=0.09$ | $p=0.77$ | AIC=1681.9 |

intervals (3a), five measures of the durational variability of voiced and voiceless intervals (3b) and two measures of the durational variability of intervals between peaks in the amplitude envelope (3c).

Results revealed low correlation coefficients, with $|r| \leq -0.15$ for all calculated correlations. Correlation tests were not significant.

### 3.2.3. The influence of segmental cues in foreign accent identification

We divided the *timeOnly* data into one subset that contained responses to the stimuli featuring uvular /ʁ/s and one subset where uvular /ʁ/s were absent (Bonferroni-adjustment: 0.05/2 = 0.025). The latter subset contained either no /ʁ/ or vocalized /ʁ/s. Fig. 7 shows boxplots of $B''_D$ and *A'* as a function of the presence or absence of uvular /ʁ/s in the stimuli. One-sample t-tests showed

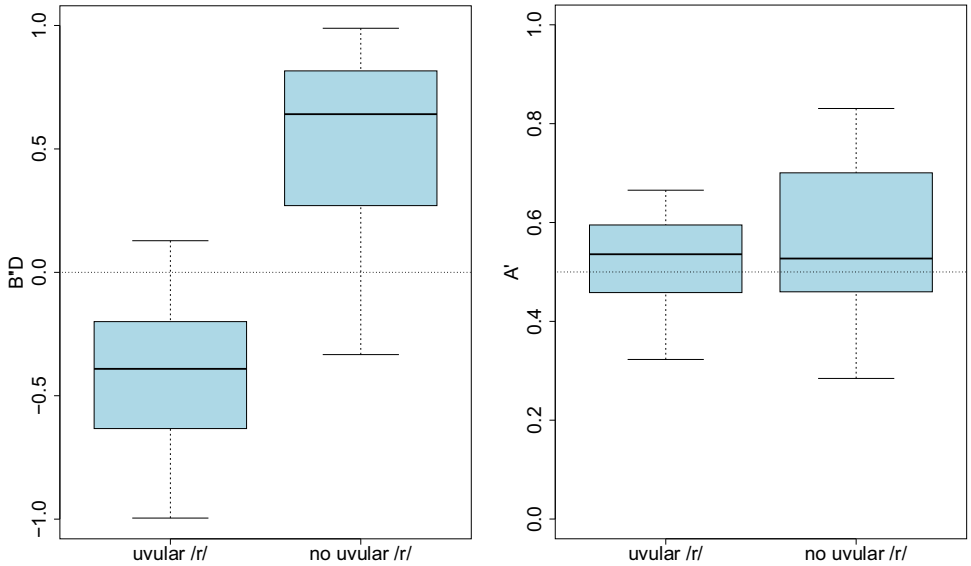



**Fig. 7.** Boxplots of listeners' accent identification performance as measured by $B''_D$ (left graph) and listeners' response bias as measured by *A'* (right graph) for stimuli of *timeOnly* which contain uvular /r/s (left boxplot) and which do not contain uvular /r/s (right boxplot). The dotted lines indicate performance at chance level and no bias, respectively.

that listeners were biased towards the response 'French' when the stimuli featured uvular /r/s ($t(19) = -5.82$, $p < 0.001^*$; $M = -0.40$, $SD = 0.31$), and that they were inclined to answer 'English' when no uvular /r/ was present in the stimuli ($t(19) = 6.31$, $p < 0.001^*$; $M = 0.52$, $SD = 0.37$). The two subsets significantly differed in bias, as measured by $B''_D$ (LM: $F(1,38) = 73.50$, $p < 0.001^*$). However, listeners' accent recognition performance, as measured by *A'*, did not differ between the two subsets (LM: $F(1,38) = 0.68$, $p = 0.41$).

## 4. Discussion

In the present paper, we reported evidence (a) that listeners can, to some extent, identify French- and English-accented German based on temporal features or on degraded spectral features alone, (b) that the combined presence of temporal and spectral cues yields an additive trend towards higher accent identification rates, and (c) that listeners' response behavior was biased depending on whether or not stimuli featured uvular /r/s. In the following, we discuss the results obtained in more detail and elaborate on potential implications for forensic phonetics and second language acquisition.

### *4.1. The importance of temporal and spectral cues in foreign accent identification*

#### *4.1.1. Temporal cues*

We found that Swiss German listeners could identify speakers' origin in French- and English-accented German based on temporal information alone in the signal condition *timeOnly*, where segment durations of foreign-accented speech were transplanted to native German speech. In previous experiments, listeners were shown to respond at chance when presented with foreign-accented stimuli that featured temporal information alone, e.g. in monotonized *sasasa*-speech (Kolly and Dellwo, 2014). However, *sasasa*-speech does not occur in natural situations, whereas the natural speech with manipulated durations used in the present experiment – albeit containing some artifacts – is assumed to sound rather familiar to listeners. This may have enhanced their identification performance, as listeners are used to interpreting temporal information in natural speech, from their everyday life. This is not the case for *sasasa*-speech.

Kolly et al. (2014) had monotonized and lowpass-filtered (<300 Hz) the sentences used for the present experiment and obtained higher accent identification rates than the ones obtained here. On the one hand, lowpass-filtered speech below 300 Hz may still have contained certain segmental cues that boosted listeners' identification performance. On the other hand, lowpass-filtered stimuli contained voicing temporal cues. Our results on the temporal patterns contained in the stimuli revealed a descriptive (but non-significant) difference between natural French- and English-accented German in voicing temporal patterns. However, voicing temporal patterns of both accents became more similar when duration transplantation was applied to create our stimuli (cf. Section 3.2.1). Next to segmental cues below 300 Hz, the additional voicing temporal cues contained in lowpass-filtered speech may therefore account for the different identification performance between the listeners tested by Kolly et al. (2014) and those tested in the present experiment, as it was previously demonstrated that the voicedness of consonants is an important cue in foreign accent identification (Flege and Port, 1981; Vieru et al., 2011). Lowpass-filtered speech also contained the original intensity features of the foreign accents at hand; however, at least their timing should have been very similar between lowpass-filtered and *timeOnly* speech, as duration transplantation was shown not to affect our amplitude peak durational measures (cf. Section 3.2.1).

Which acoustic correlates may account for listeners' sensitivity to temporal cues? We found that accent identification performance did not correlate with any of the applied acoustic measures of temporal variability. Some of these acoustic measures were shown to be affected by duration transplantation, which eliminated (descriptive) differences between French- and English-accented German (cf. Section 3.2.1.1). Considering the low overall correlations, we assume that listeners' response behavior was driven by patterns of temporal variability not revealed by the temporal measures applied in this study. For example, it may be interesting to investigate patterns of utterance-final lengthening in the future. These have been shown to differ between native and non-native accents of English (White et al., 2012), and to predict adults' and infant's discrimination of accents (White et al., 2012, 2014).

Compared to *A'*-values of 1 for perfect sensitivity, the *A'*-values reported here ($M = 0.54$, $SD = 0.07$) are fairly low. This may be due,

to some extent, to the bias driven by the /r/-variants present in the stimuli (cf. Section 3.2.3) and to some of the artifacts contained in our stimuli, which resulted from stretching certain segments and which are likely to be irritating for listeners. However, the sensitivity values reported here are in line with other experiments that use manipulated speech: Ramus et al. (2003), for example, reported mean *A'*-values between 0.57 and 0.74 for listeners' discrimination of languages based on speech temporal cues (undoubtedly, other cues come into play in accent and language identification).

#### 4.1.2. Spectral cues

Degraded spectral features of 6-band noise-vocoded speech were shown to carry enough information for listeners to identify French- and English-accented German above chance, when temporal cues were absent due to duration transplantation for the signal condition *freqOnly* (the absence of temporal cues was demonstrated in Section 3.2.1.2). This is in line with findings by Munro, et al. (2010), where listeners could identify native vs. non-native speech in utterances that were played backwards, which also largely disrupts temporal information. These findings emphasize the power of spectral information: Even when speech is strongly degraded in the frequency domain, listeners can process the remaining information, for instance the quality of certain segments, in order to identify foreign accents.

#### 4.1.3. Comparison between results based on temporal and on spectral cues

Listeners' sensitivity to reduced spectral cues in the signal condition *freqOnly* was higher than listeners' sensitivity to segmental temporal cues in *timeOnly*; this, again, emphasizes the prevalence of spectral cues for accent identification tasks.

For both signal conditions, French-accented German was identified with higher performance than English-accented German. This finding is in line with findings by Kolly and Dellwo (2014) and Kolly et al. (2014) for different types of signal-degraded speech containing primarily temporal cues. On the one hand, this may be explained to some extent by the observed tendency for listeners to be biased towards the response 'French'; bias could not be eliminated when calculating the identification performance for each accent separately (*%correct* instead of *A'*). However, there was no significant bias towards 'French' in the *timeOnly* condition. We conclude that temporal patterns of French-accented German may have sounded more salient to our listeners than those of English-accented German. This corroborates suggestions brought forth by studies in the speech rhythm domain: English and German seem to be perceptually more similar in their rhythmic organization, and they differ from French in this regard (Abercrombie, 1967; Dellwo et al., 2007; Grabe and Low, 2002; Pike, 1945; Ramus et al., 1999). Furthermore, this suggests that features of such language-specific temporal patterns are carried over to non-native speech (Arslan and Hansen, 1997; McAllister et al., 2002). Support for this idea was also reported in research by Ordin and Polyanskaya (2015), who found German learners of English to be more successful in acquiring target-like patterns of durational variability than French learners of English.

We found an overall effect of speaker, where some speakers' linguistic origin was identified with higher performance than others'. Non-native speakers thus seem to use different timing strategies when speaking a second language. Temporal features are also known to differ between speakers in their native language (Arvaniti, 2012; Dellwo et al., 2015; Leemann et al., 2014; Wiget et al., 2010). Possibly, speakers' non-native speech may be characterized by similar speaker-idiosyncratic temporal patterns as their native speech, as shown by Kolly, et al. (2015) for durational features of silent pauses. Furthermore and interestingly, accent identification scores for each speaker did not correlate with speakers' strength of foreign accent for either signal condition. This may suggest that the information retained in our *timeOnly* and *freqOnly* stimuli was not particularly salient in terms of strength of foreign accent, when listeners judged natural speech. Other features of foreign-accented speech seem to be more important for listeners' perception of accent strength.

### 4.2. The additivity of temporal and rudimentary spectral cues in foreign accent identification

The combined presence of cues from *timeOnly* as well as *freqOnly* signals in the *time+freq* condition, which contained temporal as well as degraded spectral cues in 6-band noise-vocoded speech, showed a trend towards higher accent identification performance than each type of cue separately. A significant difference was observed between performance in the *timeOnly* vs. *time+freq* condition. The finding is intuitively sound: when the information available to listeners increases, identification performance increases. This is evidence for an additive effect of temporal and spectral cues; however, the combined effect of temporal and spectral cues was smaller than the sum of single effects (Du et al., 2011; Hjalmarsson, 2011). In a similar way, Cunningham-Andersson and Engstrand (1989) have shown that perceived strength of foreign accent increases with the number of target-deviant features. We conclude that the combination of temporal and spectral cues is helpful for listeners to identify foreign accents, but it is not necessary – as each type of cue allowed accent identification above chance on its own. This is also in line with findings by Cunningham-Andersson and Engstrand (1989): some target-deviant features are more strongly associated with the perception of foreign accent than others, and different combinations of such features may increase the perception of accent strength to different degrees.

### 4.3. The influence of segmental cues in foreign accent identification

We found a significant bias depending on whether or not stimuli featured uvular /r/s. Listeners were biased towards the response 'French' when *timeOnly* stimuli featured uvular /r/s and towards 'English' when they did not. In the *timeOnly* experiment, listeners heard native German segments with French- or English-accented segment durations. However, they were not aware that the segmental content of stimuli was native German; they were only told that they would hear French- and English-accented German. All our Zurich German speakers used uvular /r/ sounds ([ʁ] or [ʁ]), and vocalized /r/s ([ɐ]): the same /r/ sounds as the ones used in Standard German from Germany. But – for the uvular /r/s – these are also the /r/ sounds used in French.

It thus seems that the listeners took the articulation of /r/ as a cue, in a task that was designed to be completed based on durational characteristics alone. Therefore, this affected their response behavior – and bias – without affecting their accent identification performance. This finding suggests that the duration transplantation method has some pitfalls when used in an identification task design, which is probably less the case when used in an experiment designed to elicit responses on accent strength (Quené and van Delft, 2010; Tajima et al., 1997; Winters and O'Brien, 2013). The finding further stresses the importance of segmental information in foreign accent identification tasks (Boula de Mareüil et al., 2008; Cunningham-Andersson and Engstrand, 1989; Vieru et al., 2011). The articulation of /r/, in particular, seems to be a crucial cue for accent identification in different target languages: Vieru et al. (2011) report it to be one of the most important cues for perceptual foreign accent identification as well as for automatic accent classification. Cunningham-Andersson and Engstrand (1989) found that target-deviant features related to the articulation of /r/ were among the ones that listeners perceived as most accented, whereas

target-deviant durational characteristics were amongst the least noticeable. Flege (1984) also cites /r/ as being a strong cue for the detection of (non-)nativeness.

### 4.4. Possible implications of this work

On the one hand, implications of this research may be found in the domain of forensic phonetics (cf. Section 1): First, the identification of a foreign accent helps narrowing down a group of suspects in forensic casework (speaker profiling; Ellis, 1994; Köster et al., 2012). Since incriminating recordings are most often made over a telephone – the quality of which cannot be controlled for –, temporal features are highly relevant. Second, foreign accent identification is relevant to some LADO cases (Cambier-Langeveld, 2010: 73; Language and National Origin Group, 2004; Verrips, 2011: 137). In LADO, telephone speech is also frequently used (Baltisberger and Hubbuch, 2010). Telephone conditions are one of the reasons for investigating listeners' accent identification performance in speech that contains temporal cues only or reduced spectral cues in general – and therefore for investigating additive effects of temporal and spectral cues in perceptual foreign accent identification.

On the other hand, this research may have implications for the domain of second language acquisition. Speakers who are discriminated against because of their particular accent and origin (Lippi-Green, 1997: 229; Schairer, 1992), for example, might wish to reduce their foreign accent to sound more native-like. It may therefore be helpful to know which accent-specific features are perceptually salient to native listeners. The present experiments suggest that French and English learners of German could take heed of temporal patterns, complementing their regular pronunciation training. Van Santen and Shih (2000) showed that durations of suprasegmental units such as the syllable strongly depend on intrinsic durations of the segments they contain. Therefore, production training focusing on the target-like pronunciation of individual segments, including their durations, may not only improve non-native speakers' production of segmental temporal patterns (e.g. vowel quantity, which is a distinctive feature of German), it could also influence the overall suprasegmental temporal features of their non-native speech towards more native-like productions (Quené and van Delft, 2010; Tajima et al., 1997). Furthermore, the pronunciation of /r/ seems to be a feature worth focusing on if a foreign accent is to be reduced.

## 5. Summary and conclusion

Our findings showed that listeners could, to a certain extent, identify the linguistic origin of French and English speakers in foreign-accented German, based solely on temporal features of these accents. Furthermore, listeners could also identify the accents in question in stimuli that contain strongly degraded spectral features alone. The combined presence of temporal and spectral information is thus not necessary for listeners to identify foreign accents better than chance. However, we found an additive trend when temporal and spectral cues were combined.

We further found that the segmental information available to listeners biased their response behavior. When stimuli featured uvular /r/s, listeners were biased towards perceiving a French accent, and a bias towards an English accent was observed in stimuli that featured vocalized or no /r/s. Segmental information – or spectral information – is highly salient and may supress listeners' attention to temporal cues to some extent. Furthermore, the /r/ pronunciation seems to be a very strong cue for listeners to make decisions about a speaker's linguistic origin. However, we found a wide range of acoustic temporal measures not to correlate with listeners' response behavior. In future work, other measures of temporal variability will have to be explored in order to explain the perceptual results presented here.

The findings may be relevant for forensic phonetics, where particular cues of foreign-accented speech allow practitioners or earwitnesses to identify a speaker's linguistic origin – and where advice often has to be given based on speech that is degraded by telephone networks or background noise. Our findings may also have implications for second language acquisition. Some nonnative speakers may wish to reduce their foreign accent. In such cases, it is crucial to know which features of an accent are perceptually salient to native listeners.

## Acknowledgments

This research was supported by the Swiss National Science Foundation (SNSF; grant numbers P1ZHP1_155024 and 100015_135287). The authors would like to thank their subjects, speakers as well as listeners, for their contribution to these experiments. Furthermore, they thank Stephan Schmid for his expert advice on foreign-accented speech, Camilla Bernardasci for her help with some of the recordings and Andrea Fröhlich for carrying out the perception experiment with some of the listeners. The present paper benefited from the helpful comments of two anonymous reviewers and the section editor, whom we would also like to thank.

## Appendix. Reading materials

01 Die Frau des Apothekers weiss immer, was sie will.
02 Das Theater hat viele neue Aufführungen geplant.
03 Er wollte sich seiner Schwächen einfach nicht bewusst werden.
04 Der öffentliche Verkehr lässt viel zu wünschen übrig.
05 Die schlechte Zahlungsbilanz lässt mich nicht zur Ruhe kommen.
06 Die Eltern geben ihm keine finanzielle Unterstützung.
07 Der starke Frühlingsregen hat grossen Schaden angerichtet.
08 Der schnellste Zug ist immer noch der ICE.
09 Der Wiederaufbau der Stadt wird sehr lange dauern.
10 Das Bildungsministerium hat den einfachsten Weg gewählt.
11 Diese Konditorei macht ausgezeichnete Kuchen.
12 Dieses Geschäft bietet sehr preisgünstige Ware an.
13 Sie haben die Wahrheit erst entdeckt, als er auspackte.
14 Für meine Mannschaft wird der Sieg ein Kinderspiel sein.
15 Die Meinungsumfragen sagen einen Sieg der Rechten voraus.
16 Die Strassen der Innenstadt wurden von der Polizei gesperrt.
17 Ein berühmtes Bild wurde aus dem Kunsthaus gestohlen.
18 Der Müssiggang ist bekanntlich aller Laster Anfang.

## References

Abercrombie, D., 1967. Elements of General Phonetics. Edinburgh University Press, Edinburgh.
Arslan, L.M., Hansen, J.H., 1997. A study of temporal features and frequency characteristics in American English foreign accent. J. Acoust. Soc. Am. 102 (1), 28–40.
Arvaniti, A., 2012. The usefulness of metrics in the quantification of speech rhythm. J. Phonet. 40, 351–371.
Auer, P., 2001. Silben- und akzentzählende Sprachen. In: Haspelmath, M., König, E., Oesterreicher, W. (Eds.). In: Language Typology and Language Universals, Vol. 2. An international handbook, Berlin/New York, de Gruyter, pp. 1391–1399.
Baltisberger, E., Hubbuch, P., 2010. LADO with specialized linguists – The development of LINGUA's working method. In: Zwaan, K., Verrips, M., Muysken, P. (Eds.), Language and Origin: The Role of Language in European Asylum Procedures. Wolf Legal Publishers, Nijmegen, pp. 9–19.
Bates, D. M., and Maechler, M. (2009). lme4: linear mixed-effects models using S4 classes. R package version 1.1-7.
Boersma, P., Weenink, D., 2014. Praat: doing phonetics by computer Version 5.4. http://www.praat.org/.
Bosker, H.R., Quené, H., Sanders, T., Jong, N.H., 2014. The perception of fluency in native and nonnative speech. Lang. Learn. 64, 579–614.

Boula de Mareüil, P., Brahimi, B., Gendrot, C., 2004a. Role of segmental and suprasegmental cues in the perception of Maghrebian-accented French. In: Proceedings of the International Conference on Spoken Language Processing 2004. Jeju, pp. 341–344.
Boula de Mareüil, P., Marotta, G., Adda-Decker, M., 2004b. Contribution of prosody to the perception of Spanish/Italian accents. In: Proceedings of Speech Prosody 2004. Nara.
Boula de Mareüil, P.and, Vieru-Dimulescu, B., 2006. The contribution of prosody to the perception of foreign accent. Phonetica 63 (4), 247–267.
Boula de Mareüil, P., Vieru-Dimulescu, B., Woehrling, C., Adda-Decker, M., 2008. Accents étrangers et régionaux en français. Traitement Autom. Lang. 49 (3), 135–163.
Byrne, C., Foulkes, P., 2004. The 'mobile phone effect' on vowel formants. J. Speech, Lang. Law 11 (1), 83–102.
Cambier-Langeveld, T., 2010. The role of linguists and native speakers in language analysis for the determination of speaker origin. J. Speech, Lang. Law 17 (1), 67–93.
Chen, B., Zhu, Q., Morgan, N., 2005. Long-term temporal features for conversational speech recognition. In: Bengio, S., Bourlard, H. (Eds.), Machine learning for Multimodal Interaction. Springer, Berlin/Heidelberg/New York, pp. 232–242.
Council of Europe, (2013). Common European framework of reference for languages: learning, teaching, assessment. http://www.coe.int/t/dg4/linguistic/source/framework_en.pdf (accessed 12.10.2015).
Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. J. Acoust. Soc. Am. 111, 2862–2873.
Cunningham-Andersson, U., Engstrand, O., 1989. Perceived strength and identity of foreign accent in Swedish. Phonetica 46, 138–154.
Dauer, R.M., 1983. Stress-timing and syllable-timing reanalyzed. J. Phonet. 11, 51–62.
Davis, M.H., Johnsrude, I.S., Hervais-Adelman, A., Taylor, K., McGettigan, C., 2005. Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. J. Exp. Psychol. 134 (2), 222–241.
Delattre, P., 1966. A comparison of syllable length conditioning among languages. Int. Rev. Appl. Linguist. Lang. Teach. 4 (3), 183–198.
Dellwo, V., 2006. Rhythm and speech rate: a variation coefficient for DeltaC. In: Karnowski, P., Szigeti, I. (Eds.), Language and Language-Processing. Frankfurt am Main, Lang, pp. 231–241.
Dellwo, V., 2008. The role of speech rate in perceiving speech rhythm. In: Proceedings of Speech Prosody 2008. Campinas, pp. 375–378.
Dellwo, V., 2010. Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence PhD Thesis. University of Bonn.
Dellwo, V., Fourcin, A., Abberton, E., 2007. Rhythmical classification of languages based on voice parameters. In: Proceedings of the International Congress of Phonetic Sciences 2007. Saarbrücken, pp. 1129–1132.
Dellwo, V., Leemann, A., Kolly, M.-J., 2012. Speaker idiosyncratic rhythmic features in the speech signal. In: Proceedings of Interspeech 2012. Portland, pp. 1584–1587.
Dellwo, V., Leemann, A., Kolly, M.-J., 2015. Rhythmic variability between speakers: articulatory, prosodic, and linguistic factors. J. Acoust. Soc. Am. 137 (3), 1513–1528.
Donaldson, W., 1992. Measuring recognition memory. J. Exp. Psychol. Gen. 121 (3), 275–277.
Du, Y., He, Y., Ross, B., Bardouille, T., Wu, X., Li, L., Alain, C., 2011. Human auditory cortex activity shows additive effects of spectral and spatial cues during speech segregation. Cereb. Cortex. 21 (3), 698–707.
Ellis, S., 1994. The Yorkshire Ripper enquiry: part 1. Forensic Linguist. 1, 197–206.
Erziehungsdirektion des Kantons Bern, (2009). Sprachniveau an der Maturität gemäss Europäischem Sprachenportfolio (ESP). http://www.erz.be.ch/erz/de/index/mittelschule/mittelschule/publikationen.assetref/dam/documents/ERZ/MBA/de/AMS/ams_sprachniveau_maturitaet.pdf, accessed 05.05.2016).
Fant, G., Kruckenberg, A., Nord, L., 1991. Durational correlates of stress in Swedish, French and English. J. Phonet. 19 (3–4), 351–365.
Ferguson, C.A., 1959. Diglossia. Word 15, 325–340.
Ferragne, E., Pellegrino, F., 2004. Rhythm in read british English: Iinterdialect variability. In: Proceedings of the International Conference on Spoken Language Processing 2004. Jeju, pp. 1573–1576.
Flege, J.E., 1984. The detection of French accent by American listeners. J. Acoust. Soc. Am. 76 (3), 692–707.
Flege, J.E., Port, R., 1981. Cross-language phonetic interference: arabic to English. Lang. Speech 24 (2), 125–146.
Grabe, E., Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. In: Gussenhoven, C., Warner, N. (Eds.), Laboratory Phonology. Mouton de Gruyter, Berlin/New York, pp. 515–545.
Green, D.M., Swets, J.A., 1966. Signal detection theory and psychophysics. Wiley, New York.
Grenon, I., White, L., 2008. Acquiring rhythm. A comparison of L1 and L2 speakers of Canadian English and Japanese. In: Proceedings of the Boston University Conference on Language Development 2008. Boston, pp. 155–166.
Hazan, V.L., Boulakia, G., 1993. Perception and production of a voicing contrast by French-English bilinguals. Lang. Speech 36 (1), 17–38.
Hirson, A., French, P., Howard, D., 1995. Speech fundamental frequency over the telephone and face-to-face: some implications for forensic phonetics. In: Windsor Lewis, J. (Ed.), Studies in general and English phonetics in honour of Professor J.D. O'Connor. Routledge, London, pp. 230–240.
Hjalmarsson, A., 2011. The additive effect of turn-taking cues in human and synthetic voice. Speech Commun. 53 (1), 23–35.
Holm, S., 2008. Intonational and durational contributions to the perception of foreign-accented Norwegian. An experimental phonetic investigation PhD Thesis. Norwegian University of Science and Technology.
Hove, I., 2002. Die Aussprache der Standardsprache in der deutschen Schweiz. de Gruyter, Berlin/New York.
de Jong, N.H., Groenhout, R., Schoonen, R., Hulstijn, J.H., 2013. Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behaviour. Appl. Psycholinguist. 34, 1–21.
Kliegl, R., Wei, P., Dambacher, M., Yan, M., Zhou, X., 2011. Experimental effects and individual differences in linear mixed models: estimating the relationship between spatial, object, and attraction effects in visual attention. Front. Psychol. 1, 1–12.
Kohler, K., 1990. German. J. Int. Phonet. Assoc. 20, 48–50.
Kolde, G. (1981). Sprachkontakte in gemischtsprachigen Städten. Vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschie- densprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i. Ue. Wiesbaden, Steiner.
Kolly, M.-J., Dellwo, V., 2014. Cues to linguistic origin: the contribution of speech temporal information to foreign accent recognition. J. Phonetic. 42, 12–23.
Kolly, M.-J., Leemann, A., Dellwo, V., 2014. Foreign accent recognition based on temporal information contained in lowpass-filtered speech. In: Proceedings of Interspeech 2014. Singapore, pp. 2175–2179.
Kolly, M.-J., Leemann, A., Boula de Mareüil, P., Dellwo, V., 2015. Speaker-idiosyncrasy in pausing behavior: evidence from a cross-linguistic study. In: Proceedings of the International Congress of Phonetic Sciences 2015. Glasgow.
Köster, O., Kehrein, R., Masthoff, K., Boubaker, Y.H., 2012. The tell-tale accent: identification of regionally marked speech in German telephone conversations by forensic phoneticians. J. Speech, Lang. Law 19 (1), 51–71.
Künzel, H.J., 2001. Beware of the 'telephone effect'. The influence of telephone transmission on the measurement of formant frequencies. Forensic Linguist. 8 (1), 80–99.
Künzel, H.J., 1997. Some general phonetic and forensic aspects of speaking tempo. J. Speech Lang. Law 4, 48–83.
Laeufer, C., 1992. Patterns of voicing-conditioned vowel duration in French and English. J. Phonet. 20 (4), 411–440.
Language and National Origin Group, 2004. Guidelines for the use of language analysis for the determination of the origin of asylum seekers. J. Speech, Lang. Law 16 (1), 113–138.
Leemann, A., 2011. Einfluss der Schweizerdeutschen Phonologie auf die Stimmhaftigkeit von Frikativen im L2-Englischen. Poster presented at the 'Phonetik und Phonologie' Conference 2011.
Leemann, A., Dellwo, V., Kolly, M.-J., Schmid, S., 2012. Rhythmic variability in Swiss German dialects. In: Proceedings of Speech Prosody 2012. Shanghai, pp. 607–610.
Leemann, A., Kolly, M.-J., Dellwo, V., 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. Forensic. Sci. Int. 238, 59–67.
Lippi-Green, R., 1997. English with an Accent: Language Ideology and Discrimination in the United States. Routledge, London/New York.
Lloyd James, A., 1929. Historical Introduction to French Phonetics. University of London Press, London.
Maassen, B., Povel, D.-J., 1985. The effect of segmental and suprasegmental corrections on the intelligibility of deaf speech. J. Acoust. Soc. Am. 78 (3), 877–886.
McAllister, R., Flege, J.E., Piske, T., 2002. The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. J. Phonet. 30 (2), 229–258.
Mermelstein, P., 1975. Automatic segmentation of speech into syllabic units. J. Acoust. Soc. Am. 58 (4), 880–883.
Munro, M.J., Derwing, T.M., Burgess, C.S., 2010. Detection of nonnative speaker status from content-masked speech. Speech Commun. 52 (7), 626–637.
Nazzi, T., Bertoncini, J., Mehler, J., 1998. Language discrimination by newborns: toward an understanding of the role of rhythm. J. Exp. Psychol. Hum. Percept. Perform. 24 (3), 756–766.
Neuhauser, S., 2011. Foreign accent imitation and variation of VOT and voicing in plosives. In: Proceedings of the International Congress of Phonetic Sciences 2003. Barcelona, pp. 1462–1465.
Ordin, M., Polyanskaya, L., 2015. Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. J. Acoust. Soc. Am. 138, 533–544.
Osberger, M.J., Levitt, H., 1979. The effect of timing errors on the intelligibility of deaf children's speech. J. Acoust. Soc. Am. 66 (5), 1316–1324.
Pettorino, M., Vitale, M., 2012. Transplanting native prosody into second language speech. In: Busà, M.G., Stella, A. (Eds.), Methodological Perspectives on Second Language Prosody. Libraria Editrice Università di Padova, Padova, Coop, pp. 11–16.
Pike, K., 1945. The Intonation of American English. University of Michigan Press, Ann Arbor.
Pinet, M., Iverson, P., 2010. Talker-listener accent interactions in speech-in-noise recognition: effects of prosodic manipulation as a function of language experience. J. Acoust. Soc. Am. 128 (3), 1357–1365.
Quené, H., van Delft, L.E., 2010. Non-native durational patterns decrease speech intelligibility. Speech Commun. 52, 911–918.

Ramus, F., Mehler, J., 1999. Language identification with suprasegmental cues: a study based on speech resynthesis. J. Acoust. Soc. Am. 105 (1), 512–521.
Ramus, F., Nespor, M., Mehler, J., 1999. Correlates of linguistic rhythm in the speech signal. Cognition 73, 265–292.
Ramus, F., Dupoux, E., Mehler, J., 2003. The psychological reality of rhythm classes: perceptual studies. In: Proceedings of the International Congress of Phonetic Sciences 2003. Barcelona, pp. 337–342.
Core Team, R. (2013). R. A language and environment for statistical computing. Version 3.0.1. Vienna. http://www.R-project.org.
Rognoni, L., Busà, M.G., 2014. Testing the effects of segmental and suprasegmental phonetic cues in foreign accent rating: an experiment using prosody transplantation. In: Proceedings of the International Symposium on the Acquisition of Second Language Speech 2013. Montreal, pp. 547–560.
Schairer, K.E., 1992. Native speaker reaction to non-native speech. Modern Lang. J. 76 (3), 309–319.
Schmid, S., 2012. The pronunciation of voiced obstruents in L2 French: a preliminary study of Swiss German learners. Poznań Stud. Contemp. Linguist. 48, 627–659.
Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., Ekelid, M., 1995. Speech recognition with primarily temporal cues. Science 270, 303–304.
Shearme, J.N., Holmes, J.N., 1961. An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. In: Proceedings of the International Congress of Phonetic Sciences 1961. Helsinki.
Tajima, K., Port, R., Dalby, J., 1997. Effects of temporal correction on intelligibility of foreign-accented English. J. Phonet. 25 (1), 1–24.
Tiffany, W.R., 1959. Nonrandom sources of variation in vowel quality. J. Speech Hear. Res. 2, 305–317.
Tilsen, S., Arvaniti, A., 2013. Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. J. Acoust. Soc. Am. 134 (1), 628–639.
van Santen, J.P.H., Shih, C., 2000. Suprasegmental and segmental timing models in Mandarin Chinese and American English. J. Acoust. Soc. Am. 107 (2), 1012–1026.
Van Zyl, M., Hanekom, J.J., 2011. Speech perception in noise: a comparison between sentence and prosody recognition. J. Hearing Sci. 1 (2), 54–56.
Verrips, M., 2011. LADO and the pressure to draw strong conclusions. J. Speech Lang. Law 18 (1), 131–143.
Vieru, B., Boula de Mareüil, P., Adda-Decker, M., 2011. Characterisation and identification of non-native French accents. Speech Commun. 53 (3), 292–310.
Vitale, M., Boula de Mareüil, P., De Meo, M., 2014. An acoustic-perceptual approach to the prosody of Chinese and native speakers of Italian based yes/no questions. In: Proceedings of Speech Prosody 2014. Shanghai, pp. 648–652.
Werlen, I., 1980. R im Schweizerdeutschen. Z. Dialektol. Linguist. 47, 52–76.
White, L., Mattys, S.L., 2007a. Calibrating rhythm: first language and second language studies. J. Phonet. 35, 501–522.
White, L., Mattys, S.L., 2007b. Rhythmic typology and variation in first and second languages. In: Prieto, P., Mascaró, J., Solé, M.-J. (Eds.), Segmental and Prosodic Issues in Romance Phonology. Benjamins, Amsterdam/Philadelphia, pp. 237–257.
White, L., Mattys, S.L., Wiget, L., 2012. Language categorization by adults is based on sensitivity to durational cues, not rhythm class. J. Memory Lang. 66 (4), 665–679.
White, L., Floccia, C., Goslin, J., Butler, J., 2014. Utterance-final lengthening is predictive of infants' discrimination of English accents. Lang. Learn. 64 (S2), 27–44.
Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., Mattys, S.L., 2010. How stable are acoustic metrics of contrastive speech rhythm? J. Acoust. Soc. Am. 127 (3), 1559–1569.
Winters, S., O'Brien, M.G., 2013. Perceived accentedness and intelligibility. The relative contributions of f0 and duration. Speech Commun. 55, 486–507.

# 7

# Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study

This chapter contains a reprint of the paper: Kolly, M.-J., Leemann, A., Boula de Mareüil, P., Dellwo, V. (2015). Speaker-idiosyncrasy in pausing behavior: Evidence from a cross-linguistic study. *Proceedings of the International Congress of Phonetic Sciences 2015*, Glasgow.[1]

Chapters 3, 4, 5, and 6 investigated the influence of speaker origin on speech temporal characteristics. Findings revealed that listeners' foreign accent identification performance is influenced by the factor *speaker*: some speakers' accent was recognized more accurately than others'. In particular, this finding did not correlate with speakers' strength of foreign accent. We therefore assume that speakers apply individual strategies in their timing of non-native speech.

In this paper we present a cross-linguistic speech production experiment, where speaker-individuality is investigated in the time domain of non-native as well as native speech. For this, 16 Zurich German speakers read 16 sentences in their native Zurich German and 16 sentences in both their non-native French and English. For each of these sentences, we computed:

- ▷ the number of silent pauses;
- ▷ the duration of silent pauses.

f

[1]ISBN: 978-0-85261-942-1.

The main outcome reported in the present paper can be outlined as follows:

⇒ Speakers produced the most and the longest pauses in their non-native French, the fewest and the shortest pauses in their native Zurich German.

⇒ Both pausing measures varied strongly between speakers, but little within speakers: For over 50% of the speakers investigated, pause number and pause durations did not vary within speaker — regardless of whether they spoke Zurich German, French, or English.

⇒ We observed an effect of sentence on measures of pausing; longer, syntactically complex sentences tended to show more and longer pauses than shorter and syntactically less complex ones.

We concluded that speakers' pausing behavior seems to be influenced by language proficiency to some extent: speakers produced the fewest and shortest pauses in their native language and the most and the longest pauses in their non-native French, in which most Zurich German speakers can be assumed to be less experienced than in English (see Sections 2.1 and 4 of the present paper). We proposed that further research on speaker-individual non-native temporal patterns should include an increase in the number of sentences analyzed by speaker and the application of a wider range of temporal measures such as the ones described in Chapter 3. Details of this future work are given in Section 10.2.

Speakers' foreign accent strength was rated by native listeners of French and English in a perception task. As speakers' accent strength is known to be strongly influenced by speaker-individual cognitive and psychological variables, we hypothesized that this variable might be speaker-specific, too; this is investigated in Chapter 8.

# SPEAKER-IDIOSYNCRASY IN PAUSING BEHAVIOR: EVIDENCE FROM A CROSS-LINGUISTIC STUDY

Marie-José Kolly[a,b], Adrian Leemann[c], Philippe Boula de Mareüil[a], Volker Dellwo[b]

[a]LIMSI-CNRS Orsay, [b]Phonetics Laboratory, Department of Comparative Linguistics, University of Zurich, [c]Phonetics Laboratory, Department of Theoretical and Applied Linguistics, University of Cambridge
{marie-jose.kolly|volker.dellwo}@uzh.ch, al764@cam.ac.uk, philippe.boula.de.mareuil@limsi.fr

## ABSTRACT

Phoneticians study acoustic speech signals. But what about the aspects of speech where the signal is silent? The present study investigated speakers' pausing behavior in their native and non-native speech. Pausing measures were applied in order to study between-speaker and within-speaker variability, where within-speaker variability was introduced by recording speakers in their native Zurich German, and in their second languages English and French. Results showed that pausing measures in the form of pause numbers and pause durations are speaker-specific. Furthermore, this speaker-specificity became evident across different languages. Results are discussed in the context of forensic voice comparison.

**Keywords**: pausing, temporal features, speaker-idiosyncrasy, second language, forensic phonetics

## 1. INTRODUCTION

Speakers, native and non-native, produce silent pauses when they speak or read aloud. Such pauses can occur in places where a pause is allowed by the syntactic makeup of the sentence – or elsewhere, where they may be perceived as "disfluencies" [18].

In the past, non-native speakers' pausing behavior was often investigated as a correlate of perceived fluency [4, 5] or as an indicator of second language proficiency [24, 25]. [6] note that pausing behavior also has to do with personality or style.

The experiment reported in the present paper was designed to explore speaker-specific pausing: two non-native speakers with the same language background and similar second language proficiency might have different habits or preferences regarding the frequency and duration of silent pauses in their speech – be it L1 or L2 speech. If this is the case, then pausing behavior may be an interesting measure for the domain of forensic phonetics. In typical cases of forensic voice comparison, trace material from a crime – e.g. recordings of a perpetrator of a bomb threat – is compared to acoustic comparison material – e.g. recordings of a suspect during a police interview – and used in forensic investigations. Acoustic measures that vary between speakers but are invariant within speakers, i.e. speaker-specific measures, are thus desirable for applications in forensic speaker comparison [22].

Speaker-specific behavior exists in different types of acoustic features. Research has revealed between-speaker variability in the frequency domain – in formant frequencies [19, 20] and fundamental frequency [13, 16, 21] –, and in the intensity domain [1]. Only recently has research shown speaker-idiosyncratic patterns in the time domain: [7, 8, 15–17] found suprasegmental temporal features to be speaker-specific and robust to within-speaker variability. Within-speaker variability introduced in forensic phonetic studies typically includes speaking style variability (read vs. spontaneous speech [8, 14–16]), channel variability (hifi vs. telephone speech, [14, 15]), and voice disguised speech [13].

Do speakers differ in their pausing behavior? And does pausing behavior remain speaker-specific if speakers talk in different languages? We introduced between-speaker variability by studying 16 speakers, and included within-speaker variability by having the same speakers produce native Zurich German speech, and non-native English and French speech.

## 2. METHODS

### 2.1. Speakers

16 speakers of Zurich German (eight male / eight female) were recorded at the Phonetics Laboratory, University of Zurich, to create the TEVOID corpus [7, 8, 15–17]. Speakers' age ranged between 20 and 33 years (M=25.4; SD=3.7). All speakers were University of Zurich students who spoke the dialect of the city of Zurich. They thus showed little to no regional accent variability, as attested by informal listening tests. All speakers had learned French and English as a second language at school. Usually, French classes started at age 8 and English classes at age 13. The speaker group was thus relatively homogeneous in terms of native dialect, age, education and second languages spoken. Recordings were made in a sound-treated booth using an omnidirectional Earthworks QTC40 high definition

condenser microphone (sampling rate of 44.1kHz; 16 bit quantization). Speakers were paid 30 Swiss Francs per hour for their participation.

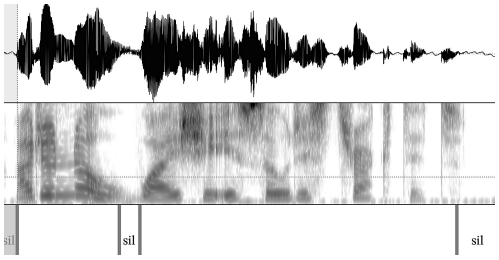
### 2.2. Material

Each speaker read 16 Zurich German sentences, 16 English sentences and 16 French sentences taken from the TEVOID corpus. English and French sentences were literal translations of the Zurich German sentences (yet idiomatic in English and French) and were thus roughly similar in length: sentences typically contained 15–20 syllables. These 768 sentences (16 speakers × 16 sentences × 3 languages) constituted the corpus used in the present study. Prior to the recording, speakers had prepared reading the sentences at home, to ensure fluent reading of the material. If hesitations in the form of filled pauses occurred in a sentence, speakers repeated the sentence spontaneously or, if not, they were asked to do so. Sentences which contained hesitations in the form of silent pauses were not repeated, however.

### 2.3. Data editing

To prepare the data for the application of pausing measures (cf. 2.4), trained phoneticians (first and second author) labeled each sentence for silent pauses using Praat software [3]. Speakers may pause to reflect syntactic constituents in spoken language, e.g. between a main and a subordinate clause, to mark conversational structure, e.g. emphasize a subsequent stretch of speech, or for stylistic reasons, e.g. to reflect idiolectal aspects of speech. In addition, speakers may pause for cognitive reasons, e.g. hesitating as a means to prepare for what to say next. All these types of silent pauses were labeled in our corpus, which means that no duration threshold was applied for the labeling of silent parts. Pauses were labeled perceptually – every silent part which was perceived as a pause was labeled as such (indicated by the interval label *sil* in Figure 1). Every sentence in the corpus is preceded and followed by a (labeled) pause, cf. Figure 1.

**Figure 1:** Praat TextGrid with hand-labeled pauses in the English sentence *I don't know* [pause] *why she is so distracted*.




### 2.4. Pausing measures applied

We applied two measures that describe speakers' pausing behavior and are widely used in second language research [4–6, 9, 10, 14, 18, 24–27]:

1. The number of pauses in a sentence: *pauseNbr*.
2. The sum of the durations (in seconds) of all pauses in a sentence: *pauseDur*.

The silences that precede and follow each sentence were not taken into account for the calculation of *pauseNbr* and *pauseDur*.

### 2.5. Speech tempo effects

Findings of [9, 26, 27] suggest that, for some speakers, pausing behavior covaries with articulation rate. We therefore checked whether *pauseNbr* or *pauseDur* may be influenced by articulation rate. As a measure of articulation rate, we calculated *ratePeak*: the number of automatically detected peaks in the amplitude envelope – which roughly corresponds to the number of syllables – per second, excluding pauses [7]. Neither *pauseNbr* (r=0.16) nor *pauseDur* (r=0.10) were correlated with *ratePeak*.

### 2.6. Statistical analyses

Data were analyzed using linear mixed effect models (LMEs), with R software [23] and the R package *lme4* [2]. *Language* was included as a fixed effect, *speaker* and *sentence* as random effects. We included a random slope of *speaker* on *language* to test for interactions between the two factors. Effects were tested by model comparison between a full model in which the factor in question was present and a reduced model in which the factor was excluded. We applied standard likelihood ratio tests to compare the two models. We report AIC (Akaike Information Criterion) values for the relative goodness of fit [12]. We assumed an $\alpha$ level of 0.01.

## 3. RESULTS

### 3.1. Number of pauses: *pauseNbr*

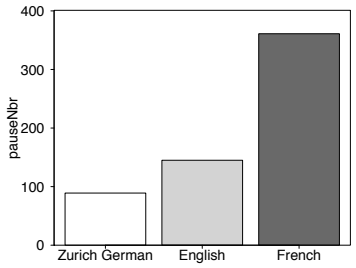
Table 1 summarizes the results obtained for *pauseNbr*. The AIC values are equal for each test because they are based on the full model, which, for every factor, provided an improved goodness of fit.

**Table 1**: Summary of the LMEs for *pauseNbr*

| Factor | Result |
|---|---|
| *language* | p<0.0001; AIC=1740 |
| *speaker* | p<0.0001; AIC=1740 |
| *language*speaker* | p<0.0001; AIC=1740 |
| *sentence* | p<0.0001; AIC=1740 |

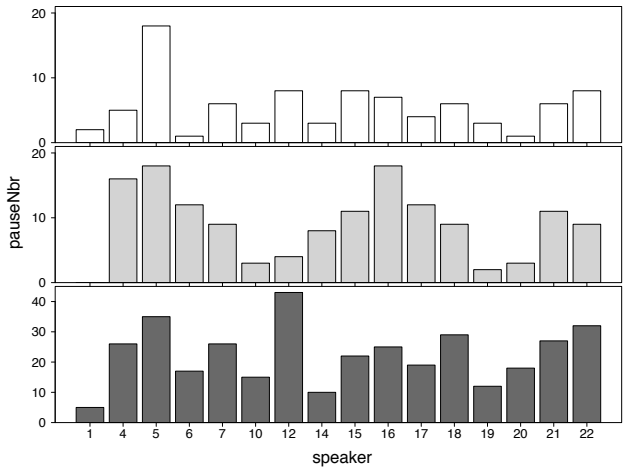
*language* was found to be highly significant, cf. Figure 2: *pauseNbr* was lowest in Zurich German (M=0.35, SD=0.63), followed by English (M=0.57, SD=0.69) and French (M=1.41, SD=1.22).

**Figure 2**: Barplots of *pauseNbr* per *language* for Zurich German, English, and French.




We also found a highly significant effect of *speaker*, cf. Figure 3. Since there was a significant interaction between *language* and *speaker*, we calculated simple effects for the factor *speaker* on the Zurich German, English and French data separately (Bonferroni corrected α: 0.01/3=0.003). *speaker* was significant in the Zurich German ($p<0.0001$; AIC=431), English ($p<0.0001$; AIC=508) as well as in the French ($p<0.0001$; AIC=730) data. We also calculated simple effects for the factor *language* for each speaker separately (Bonferroni corrected α: 0.01/16=0.0006). *language* was only significant in 7 out of 16 speakers. Furthermore, *pauseNbr* was affected by the highly significant factor *sentence*.

**Figure 3**: Barplots of *pauseNbr* by *speaker* for Zurich German (top, white), English (center, light gray), and French (bottom, dark gray). NB: y-axes show different maxima.




### 3.2. Pause durations: *pauseDur*

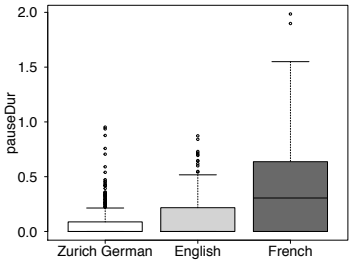
Table 2 summarizes the results obtained for *pauseDur*.

**Table 2**: Summary of the LMEs for *pauseDur*.

| Factor | Result |
|---|---|
| *language* | $p<0.0001$; AIC=-139 |
| *speaker* | $p<0.0001$; AIC=-139 |
| *language*speaker* | $p<0.0001$; AIC=-139 |
| *sentence* | $p<0.0001$; AIC=-139 |

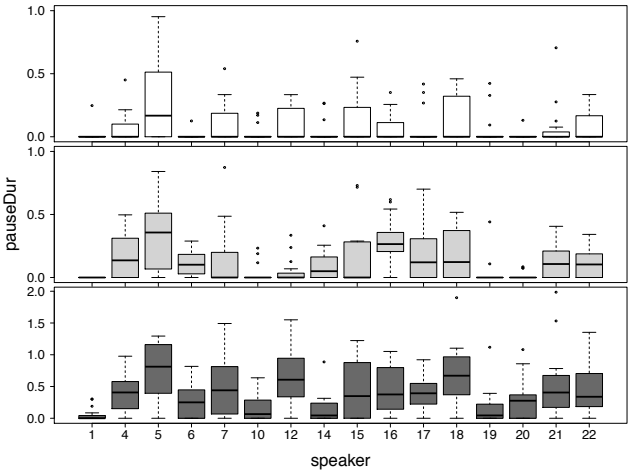
*language* was highly significant, cf. Figure 4: *pauseDur* was lowest in Zurich German (M=0.08, SD=0.17), followed by English (M=0.13, SD=0.18) and French (M=0.40, SD=0.41). There was a highly significant effect of *speaker*, cf. Figure 5. Since the interaction of *language* and *speaker* was significant, we calculated simple effects for *speaker* as described in 3.1. *speaker* was significant in Zurich German ($p<0.0001$; AIC=-244), English ($p<0.0001$; AIC=-174) as well as French ($p<0.0001$; AIC=-142).

**Figure 4**: Boxplots of *pauseDur* by *language* for Zurich German, English, and French.




We also calculated simple effects for *language* (cf. 3.1). Again, *language* was only significant in 7 out of 16 speakers. Furthermore, *pauseDur* was affected by the highly significant factor *sentence*.

**Figure 5**: Boxplots of *pauseDur* by *speaker* for Zurich German (top, white), English (center, light gray), and French (bottom, dark gray). NB: y-axes show different maxima.



## 4. DISCUSSION

In terms of *language* effects, we found that speakers produced the fewest and the shortest pauses in their native Zurich German speech, and the most and the longest pauses in their French speech. Speakers' pausing behavior in English was located in between French and German. This may be explained by the cognitive task at hand: speaking a second language is cognitively more demanding than speaking a first language – in which speakers are more proficient: [10] has shown that cognitively more demanding tasks lead to longer pauses in speech. This is corroborated by [24], who shows that second language proficiency affects the number and duration of pauses. The difference between the two second languages French and English in our data is most likely explained by the fact that Zurich German speakers are more proficient in English than in French. Even though, at school, they learned French before English and even though they live in a country where French is an official language, Swiss German university students most probably hear and produce English more often than French.

We found an effect of *sentence* for the number and the duration of pauses. Looking at the data more closely, results showed that certain sentences – such as English (E1): *One could either help serving in the house, or go outside* and French (F1): *Soit on aidait à servir là-bas, dans la maison, soit on allait dehors* – show many and long pauses. In (E2): *I am really interested in everything* and (F2): *Je suis vraiment intéressée à tout*, there were fewer and shorter pauses. (E1) and (F1) are some of the longest sentences of the corpus, and thus are more likely to show pauses because of that. Furthermore, their syntactic construction provides potential slots for pauses that co-occur with punctuation such as commas. (E2) and (F2), on the other hand, are very short and made up of one main clause only. Fewer pauses are thus expected in these sentences.

A higher number of pauses is expected to lead to more occurrences of phrase-final lengthening and thus to a lower articulation rate. This was not the case in our corpus: number and duration of pauses were not related to our measure of articulation rate. This finding may be due to the – possibly too coarse – peak detection method applied, which leaves room for further investigations in the future.

In terms of *speaker* effects, our data revealed significant between-speaker differences, in the number as well as the duration of pauses. At the same time, measures varied little within speakers: only for 7 out of 16 speakers did we observe a simple effect of *language*. Speaker 1, for example, made few pauses in Zurich German, French as well as in English speech. Speaker 5, on the other hand, showed high values for the number of pauses in all three languages. The same holds for pause durations: speaker 1 produced short pauses in all three languages, whereas speaker 5 produced long pauses.

As for the implications for the domain of forensic phonetics, both pausing measures showed significant between-speaker variability on the one hand and little within-speaker variability on the other. When testing simple effects of *language*, 7 out of 16 speakers did not differ in their pausing behavior – regardless of whether they spoke Zurich German, English or French. Furthermore, Figures 3 and 5 show that, even if there was an effect of *language* for a particular speaker, the direction of the effect was most often constant: speakers produced most and the longest pauses in French and least and the shortest pauses in Zurich German. This is surprising, since [14] found low speaker-specific values for speakers' pausing behavior, whereas within-speaker variability – introduced by having speakers read and speak spontaneously – was relatively high.

The International Association for Forensic Phonetics and Acoustics (IAFPA, [11]) advises members to "exercise particular caution" when carrying out analyses on non-native speech. More extensive research about L2 speech may complement existing parameters that are used in forensic casework. More importantly, incriminating speech samples are frequently recorded over a telephone, which degrades the spectral characteristics of the acoustic signal but does not affect temporal characteristics such as pausing [14].

## 5. CONCLUSION AND FUTURE WORK

The present study set out to investigate whether pausing behavior is speaker-specific, and the degree to which this is true across different languages. Results showed high between- and low within-speaker variability in the number and duration of pauses in each sentence. This suggests that temporal measures such as speakers' pausing behavior may be useful for the domain of forensic voice comparison. Further steps in this research will include an increase in size of the database and the application of a wider variety of temporal measures, cf. [7, 8, 15–17].

## 6. ACKNOWLEDGEMENTS

This research was supported by the Swiss National Science Foundation (SNSF; grant numbers 135287 and 155024). We would like to thank Bob Ladd for valuable feedback on a first version of this manuscript and Stephan Schmid for his expert advice on foreign-accented speech.

## 7. REFERENCES


[1] Amino, K., Arai, T. 2009. Speaker-dependent characteristics of the nasals. *Forensic Science International* 185, 21–28.

[2] Bates, D.M., Maechler, M. 2009. lme4: Linear mixed-effects models using S4 classes. R package version 1.1-7.

[3] Boersma, P., Weenink, D. 2012. *Praat: Doing phonetics by computer*. http://www.praat.org/.

[4] Bosker, H. R., Quené, H., Sanders, T., Jong, N. H. 2014. The perception of fluency in native and nonnative speech. *Language Learning* 64, 579–614.

[5] Cucchiarini, C., Strik, H., Boves, L. 2002. Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America* 111, 2862–2873.

[6] de Jong, N. H., Groenhout, R., Schoonen, R., Hulstijn, J. H. 2013. Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behaviour. *Applied Psycholinguistics* 34, 1–21.

[7] Dellwo, V., Leemann, A., Kolly, M.-J. 2012. Speaker idiosyncratic rhythmic features in the speech signal. *Proceedings of Interspeech 2012*, Portland, USA.

[8] Dellwo, V., Leemann, A., Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic and lexical factors. To appear in: *Journal of the Acoustical Society of America* 137, 1513–1528.

[9] Fougeron, C., Jun, S.-A. 1998. Rate effects on French intonation: Prosodic organization and phonetic realization. *Journal of Phonetics* 26, 45–69.

[10] Grosjean, F. 1980. Temporal variables within and between languages. In: Dechert, H. W., Raupach, M. (eds), *Towards a Cross-Linguistic Assessment of Speech Production*. Frankfurt: Lang, 39–53.

[11] IAFPA = International Association for Forensic Phonetics and Acoustics. http://www.iafpa.net.

[12] Kliegl, R., Wei, P., Dambacher, M., Yan, M., Zhou, X. 2011. Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology* 1, 1–12.

[13] Künzel, H. J. 2000. Effects of voice disguise on speaking fundamental frequency. *Forensic Linguistics* 7, 149–179.

[14] Künzel, H. J. 2013. Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech Language and the Law* 4, 48–83.

[15] Leemann, A., Kolly, M.-J., Dellwo, V. 2014. Speaker-individuality in the time domain: Implications for forensic voice comparison. *Forensic Science International* 238, 59–67.

[16] Leemann, A., Mixdorff, H., O'Reilly, M., Kolly, M.-J., Dellwo, V. (2015). Speaker-individuality in Fujisaki model f0 features: Implications for forensic voice comparison. *International Journal of Speech, Language and the Law* 21, 343–370.

[17] Leemann, A., Kolly, M.-J., Dellwo, V. (in review). Speaker-invariant suprasegmental temporal features in normal and disguised speech.

[18] Lennon, P. 1990. Investigating fluency in EFL: A quantitative approach. *Language Learning* 40, 387–417.

[19] McDougall, K. 2004. Speaker-specific formant dynamics: An experiment on Australian English /aI/. *International Journal of Speech, Language and the Law* 11, 103–130.

[20] Morrison, G. 2009. Likelihood-ration based forensic speaker comparison using representations of vowel formant trajectories. *Journal of the Acoustical Society of America* 125, 2387–2397.

[21] Nolan, F. 2002. Intonation in speaker identification: An experiment on pitch alignment features. *Forensic Lingistics* 9, 1–21.

[22] Nolan, F. 2009. *The phonetic bases of forensic speaker identification*. 2nd ed. Cambridge: Cambridge University Press.

[23] R Core Team 2013. R. A language and environment for statistical computing. Version 3.0.1. Vienna. http://www.R-project.org.

[24] Riazantseva, A. 2001. Second language proficiency and pausing. A study of Russian speakers of English. *Studies in Second Language Acquisition* 23, 497–526.

[25] Trofimovich, P., Baker, W. 2006. Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition* 28, 1–30.

[26] Trouvain, J. 2003. *Tempo variation in speech production. Implications for speech synthesis*. PhD Thesis, University of Saarbrücken.

[27] Trouvain, J., Grice, M. 1999. The effect of tempo on prosodic structure. *Proceedings of the 14th ICPhS*, San Francisco, 1067–1070.

# 8

# Strength of foreign accent is speaker-specific across different non-native languages

This chapter presents an experiment carried out to describe speakers' strength of foreign accent in the Non-native Speaker-Individuality Corpus. In Chapter 7, we reported that the speakers in this corpus varied from each other in their pausing behavior while exhibiting relatively robust pausing within speaker, across languages. Since accent strength is known to be strongly influenced by speaker-individual cognitive and psychological variables, we hypothesized that accent strength may also be speaker-specific to some extent, which is investigated in the present experiment. To characterize each speaker's accent strength, we ran two perception experiments where listeners rated speech for accent strength on a continuous scale:

- ▷ native French listeners rated French as spoken by Zurich German speakers;
- ▷ native English listeners rated English as spoken by Zurich German speakers.

The main findings of the present experiment were the following:

- ⇒ Speakers' foreign accent was perceived as stronger in their non-native French than in their non-native English speech.
- ⇒ Between-speaker variability in accent strength was significant.
- ⇒ Accent strength was mostly language-invariant within speakers.

We concluded that speakers' accent strength is influenced not only by language proficiency, but also by speaker-individual characteristics that do not vary with language spoken. The characterization of speakers' accent strength may be of use in future work, where the potential in terms of speaker-individuality of a wider range of time domain measures should be tested, using the Non-native Speaker-Individuality Corpus (see Section 10.2).

# Introduction

One can be highly proficient in a non-native language and still be recognized as a second language speaker based on just a few sounds or syllables. Non-native speakers may have a native-like grammar and lexicon, but their speech may still be characterized by an unmistakable foreign accent. When starting to acquire a second language beyond early childhood, most speakers can be perceived as second language speakers through their non-native speech (Major, 2001; Moyer, 2004). There are exceptions to this observation, however: Moyer (2004) describes a few so-called "exceptional learners" who started learning German during their adulthood and are perceived as native-like by listeners most of the time. At the other end of the scale, the presence of a foreign accent has been observed even in speakers who started acquiring their second language between five and seven years of age (Flege, 1992).

Why does a particular speaker have a foreign accent — and one that is perceived as being stronger than another speaker's foreign accent, even though both speakers' experience with the non-native language is comparable? Other than external factors such as the amount of time a speaker has spent in a place where the non-native language is predominantly spoken, the frequency with which s/he still uses the native language, the duration and intensity of language classes, and the amount of contact with native speakers of the non-native language, a range of cognitive characteristics of the speaker seem to play a role in accent strength (Major, 2001). However, even if such factors are held constant, between-speaker differences in foreign accent strength are observed. Research has shown that these differences can be explained to some degree by social-psychological variables such as speakers' attitude towards the second language, their intrinsic and extrinsic motivation for learning the language, their sense of identity towards the native language, empathy, self-esteem, or musicality (Major, 2001). In particular, empirical investigations by Moyer (2004) and Kolly (2011) report that speakers' attitude towards the second language correlates with the strength of their foreign accent.

Acoustic correlates of perceived accent strength are manifold. However, it has been shown that non-native temporal characteristics such as segment durations (Tajima et al., 1997; Holm, 2008; Quené and van Delft, 2010; Winters and O'Brien, 2013), pausing characteristics (Trofimovich and Baker, 2006), and speaking rate (Dellwo, 2010) increase the perceptual impression of foreign accent (and decrease intelligibility). In the following we present data on between-speaker and within-speaker variability in foreign accent strength. Accent strength was rated in a perceptual task in which Zurich Germans' French speech was rated by native French listeners and their English speech by native English listeners. We further correlate accent strength with two measures of pausing applied to the same Zurich Germans' non-native speech (see Chapter 7).

# Materials and methods

## Stimuli

For this experiment we used the Zurich-German-accented French and English speech material from the Non-native Speaker-Individuality corpus described in Section 2.2 and Chapter 7. For each of the 16 Zurich German speakers, a subset of 10 sentences per speaker and language was used in the accent rating task, making for a total of 160 stimuli (16 speakers × 10 sentences) per language. All stimuli were scaled to an intensity of 70 dB.

## Subjects

16 native French listeners (5 male, 11 female) from the Université de Nanterre in Paris rated the French sentences, and 16 native English listeners (7 male, 9 female) from the University of Cambridge rated the English sentences for foreign accent strength. All listeners were students at these universities. French subjects ranged in age between 19 and 29 years (M=22.31, SD=2.91) and English subjects between 18 and 33 (M=20.75, SD=4.04). None of the listeners had knowledge of German, and none of them reported significant problems with hearing or sight.

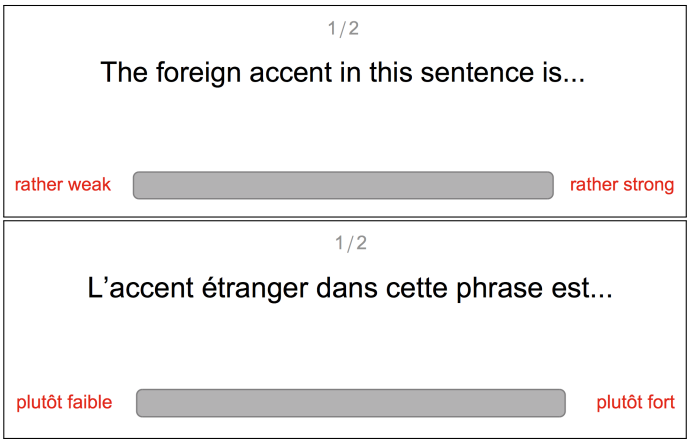



Figure 8.1: Interface for rating accent strength in the French (top) and English (bottom) experiment.

## Procedure

Subjects were tested individually in quiet rooms at the Université de Nanterre and the University of Cambridge, respectively. They were presented with the 160 stimuli over high-quality headphones. Stimulus order was randomized separately for each subject. Subjects were told that they would hear sentences in their native language spoken by

native speakers of Swiss German and that they would have to rate the intensity of the speakers' foreign accent on a continuous scale, for each sentence. They were encouraged to use the entire scale, and to respond intuitively. Before the start of the experiment, listeners were familiarized with the experiment interface and with foreign-accented speech through the presentation of two random stimuli from the experiment. Listeners responded by clicking within the quasi-continuous scale using an experiment interface (a custom-made Praat plug-in tool, see Figure 8.1) on a laptop computer. The experiment lasted between 15 and 20 minutes and listeners were paid the equivalent of 30 Swiss Francs per hour for their participation.

### Data analysis and statistical analyses

The quasi-continuous scale presented to our listeners was divided into 100 intervals. When listeners responded by clicking within the grey scale (see Figure 8.1), the number of the interval they clicked was saved; accent degree could potentially range from 1 to 100. Data were analyzed using R (R Core Team, 2014) and the R-package *lme4* for the calculation of linear mixed effects models (LME; Bates and Maechler, 2009). Our model included *speakers' gender* and *language* as fixed effects and *listener*, *speaker*, and *sentence* as random intercepts. We further included a by-*speaker* random slope on the effect of *language* to test for interactions between the two factors. Effects were tested by model comparison between a full model in which the factor in question was present and a reduced model in which the factor was excluded. We applied standard likelihood ratio tests to compare the two models (R code: anova(model_full, model_reduced)) and we report AIC (Akaike Information Criterion) values for the relative goodness of fit. We assumed an $\alpha$-level of 0.05. For correlations we indicate a non-parametric coefficient, Spearman's $r$.

## Results

Table 8.1 summarizes the results of the statistical analyses carried out on the accent strength data. We observed a significant effect for *language*, *speaker*, and *sentence*, as well as for the interaction between *speaker* and *language*. As for the effect of *language*, speakers had a stronger perceived foreign accent in French (M=53.83; SD=29.80) than in English (M=49.56; SD=25.09) speech. The effect of *speaker* reveals that between-speaker variability is high. As there was a significant interaction between *language* and *speaker*, we report simple effects. Simple effects revealed significant between-speaker variability in accent strength in non-native French ($\chi^2(1)$=605.34, AIC=23009, p<0.001*) as well as English speech ($\chi^2(1)$=734.56, AIC=22790, p<0.001*; Bonferroni-adjustment: 0.05/2=0.025). However, simple effects of language were significant only for speaker 1 ($\chi^2(1)$=14.441, AIC=2874.2, p<0.001*; Bonferroni-adjustment: 0.05/16=0.003). We therefore observed strong between-speaker variability in accent strength; at the same time, accent strength remained relatively robust within speakers, regardless of whether speakers spoke French or English. This is illustrated in Figure 8.2.

| Factor | Result | | |
|---|---|---|---|
| *language* | $\chi^2(3)$=125.37, | AIC=45907, | p<0.001* |
| *speaker* | $\chi^2(3)$=1502.70, | AIC=45907, | p<0.001* |
| *language*speaker* | $\chi^2(2)$=123.84, | AIC=45907, | p<0.001* |
| *sentence* | $\chi^2(3)$=195.70, | AIC=45907, | p<0.001* |

Table 8.1: Summary of statistics for accent strength.

Our data further revealed a high correlation between the number of pauses speakers produce and speakers' accent strength, for non-native French speech ($r$=0.92). For French, we also found a high correlation between pause durations and perceived accent strength ($r$=0.87). However, correlations between accent strength and the number of pauses ($r$=0.18) or pause durations ($r$=0.26) were low in the non-native English material.

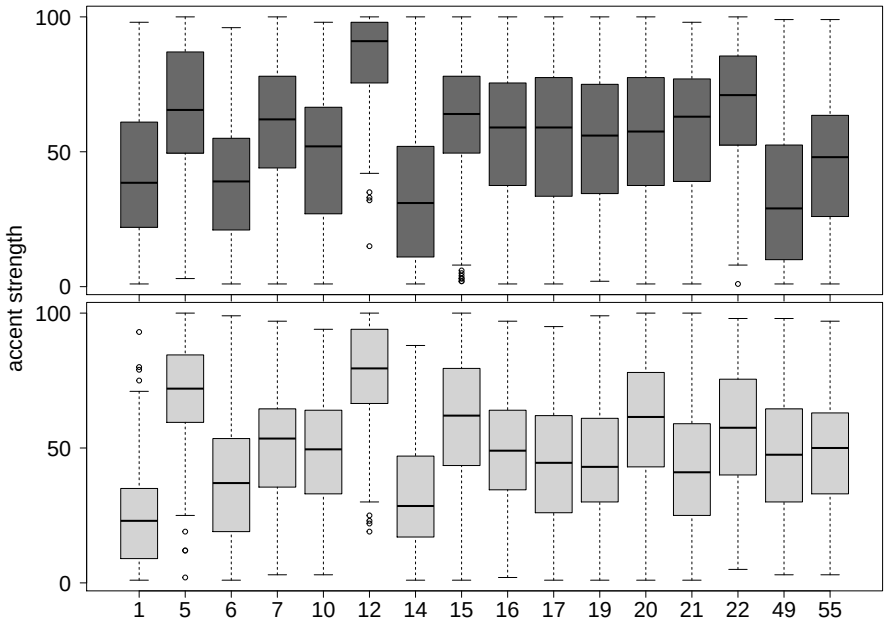



Figure 8.2: Boxplots for speakers' accent strength in French (top, dark gray) and English (bottom, light gray).

# Discussion and conclusion

Results revealed a significant effect of *language*. Overall, speakers were perceived to have a stronger accent in their French than in their English non-native speech. This may be related to the fact that Zurich German speakers are probably more proficient in English

than in French. Even though, at school, they learn French before English and even though they live in a country where French is an official language, Swiss German university students such as our speakers usually hear and produce English more often than French. In particular, these speakers were also observed to produce more and longer pauses in French speech than in English speech (see Chapter 7). An alternative explanation for the effect of *language* may be that our French listeners had a more normative model in mind when rating foreign-accented French than was the case for English listeners, who are possibly more used to variation within their native language.

We found a significant effect of *speaker* on accent strength as rated by native listeners of French and English. Interestingly — and this effect is mirrored in the speech production data in Chapter 7 — the observed between-speaker variability is complemented by within-speaker invariance: Only for one speaker did we find a significant effect of *language* when testing simple effects. Speakers 6 and 14, for example, were perceived to have a weak accent in French as well as in English. Speakers 5 and 12 were perceived to have a strong accent in both their second languages. This combined effect of high between-speaker and low within-speaker variability is desirable in the field of forensic voice comparison (Nolan, 2009). It may be that accent strength is highly speaker-individual because it is determined not only by factors external to the speaker — e.g., age of acquisition (see Major, 2001, for an overview) — but also to a considerable degree by cognitive as well as social-psychological factors, as described by Moyer (2004) and Kolly (2011).

We observed a significant effect of *sentence* in both the French and the English material. Exemplary qualitative analyses revealed that certain sentences contain a high number of potentially target-deviant features that may be perceptually salient to the listeners in terms of accent strength. For example, the French sentence *Soit on aidait à servir là-bas, dans la maison, soit on allait dehors* 'One could either help serving in the house, or go outside' exhibited the highest ratings for accent strength. It contains as many as four nasal vowels, which are known to be difficult for (Swiss German) learners of French. The sentence *Le Japon a besoin de beaucoup d'électricité, surtout en été* 'Japan needs a lot of electricity, especially in the summer' obtained low ratings of accent strength. It contains only three nasal vowels. It is conceivable that target-deviant articulations of nasal vowels, for example, were perceptually salient to native listeners and led them to give lower ratings to sentences that contained a higher number of such vowels. The second sentence is also much shorter and syntactically less complex than the first. Sentence length, syntactic complexity as well as punctuation may lead to different pausing behavior or fluency in general, which may, in turn, influence ratings of accent strength (Trofimovich and Baker, 2006).

Accent strength was highly correlated with pausing characteristics in French, but not in English non-native speech. This may be due to the overall low number of pauses in the non-native English material (see Chapter 7). It may be that pausing characteristics, if present in non-native speech, are highly salient in terms of accent. In the absence of pausing characteristics, however, listeners may focus their attention on other — segmental

and suprasegmental — cues for their accent rating. The relationship between speakers' accent strength and temporal characteristics of their non-native speech, as well as the potential of such findings for the domain of forensic phonetics will have to be explored further in the future (see Section 10.2).

# Part III

# General discussion and conclusions

# 9

# General discussion

The experimental investigations presented in Part II put forth a wide range of findings that are summarized and discussed in this chapter. Section 9.1 reviews the results related to listeners' identification of speaker origin based on temporal and rhythmic features, whereas Section 9.2 discusses the results of our experiment on language-independent speaker-individual durational features and our findings on speaker-specific accent strength across languages. Section 9.3 briefly summarizes by-products of the experiments that point to general properties of non-native speech. Finally, Section 9.4 presents possible implications of the present work.

## 9.1 Cues to speaker origin in the time domain

In a series of 10 perception experiments presented in Chapters 4, 5, and 6, we reported the results of listeners' foreign accent identification performance based on (primarily) temporal features. For each experiment we presented between 10 and 20 Swiss German university students with acoustic stimuli in a particular signal condition and — for signal conditions that rendered speech unintelligible — with the sentence transcripts corresponding to the acoustic stimulus.

Table 9.1 presents the 10 signal manipulations applied and, for each manipulation, details on the type of time domain information highlighted by it (segment durations, amplitude envelope temporal information, or temporal patterns of voicing) and the degree to which it reduces frequency domain information. Synthesizing findings presented in Chapters 4, 5, and 6, Table 9.1 also presents a ranking of signal manipulations according to listeners' accent identification performance (measured by mean values calculated for *A'*, see Chapter 6), where the signal manipulations that allowed listeners to identify French- and English-accented German above chance are labeled with a checkmark. Together with the results presented and discussed in Part II of this thesis, this allows cue weighting for listeners' identification of French- and English-accented speech.

| Signal manipulation | time domain | | | frequency domain | | |
|---|---|---|---|---|---|---|
| | segments | amplitude | voicing | present | reduced | absent |
| ✓none | + | + | + | ✓ | | |
| ✓1-bit requantization (st) | + | − | + | | ✓ | |
| ✓lowpass filter (st) | − | + | + | | ✓✓ | |
| ✓6-band noise vocoding (st) | − | + | − | | ✓ | |
| ✓*freqOnly* (st) | − | − | − | | ✓ | |
| ✓lowpass filter + monotonization (st) | − | + | + | | ✓✓✓ | |
| ✓*timeOnly* | + | − | − | | | ✓ |
| 3-band noise vocoding (st) | − | + | − | | ✓✓✓ | |
| 6-band noise vocoding | − | + | − | | ✓✓✓ | |
| *sasasa* + monotonization (st) | − | − | + | | | ✓ |

Table 9.1: Signal manipulation conditions, the type of information they contain in the time domain, and the amount of information they contain in the frequency domain. For *timeOnly*, non-native segment durations were transplanted to native speech; for *freqOnly*, native segment durations were transplanted to non-native speech that was subsequently 6-band noise vocoded; *st* refers to the presentation of sentence transcripts. Signal conditions are ranked according to listeners' accent identification performance measured by *A'*, and preceded by a checkmark in cases where they allowed accent identification above chance.

## Time domain cues

Listeners were shown to recognize French- and English-accented German above chance in the signal condition *timeOnly*, where no frequency domain information was present. This provides evidence that French- and English-accented German do in fact differ in their rhythmic or temporal organization, and that listeners can use this temporal information as a cue to speaker origin. The experiment therefore confirms that temporal patterns of non-native speech are influenced, to some extent, by interference from speakers' native language (Caramazza et al., 1973; Flege and Hillenbrand, 1984; Flege et al., 1992; Hazan and Boulakia, 1993; Arslan and Hansen, 1997; McAllister et al., 2002).

**Segment durations** The signal condition *timeOnly* contained non-native segment durations only. The fact that it allowed for accent identification above chance reveals that segment durations are, to some extent, perceptually salient in terms of speaker origin. This was also suggested by the relatively high identification performance in 1-bit requantized speech (which, additionally, contained reduced frequency domain information), as discussed in Chapter 4, Section 5. Furthermore, as the signal condition *timeOnly* contained the original segment durations of French- and English-accented German, it also featured suprasegmental temporal information of these accents, since suprasegmental temporal units are known to be strongly influenced by the intrinsic durations of the segments they contain (van Santen and Shih, 2000). Segmental as well as suprasegmental temporal information can thus be assumed to play into listeners' accent identification performance.

**Voicing temporal patterns** Monotonized *sasasa*-speech contained non-native voicing temporal patterns only. The analysis of the temporal patterns contained in unmanipulated French- and English-accented German speech, as presented in Chapter 3, suggested that the percentage over which speech is voiced, *%VO*, may be one of the strongest indicators of speaker origin within the range of temporal variables studied — more so than characteristics of vocalic and consonantal intervals, for example. Furthermore, research has shown that voicing is an important perceptual cue to foreign accent identification (Flege and Port, 1981; Vieru et al., 2011). However, our monotonized *sasasa*-speech did not allow listeners to identify speaker origin. Does this mean that voicing temporal information is perceptually less salient than the segment durations present in *timeOnly* speech, which did not feature (measurable) voicing differences between French- and English-accented German (see Chapter 6, Section 2.5)? Not necessarily, as signals such as *sasasa*-speech are never encountered in everyday listening situations. Listeners' lack of familiarity with this signal condition which contains segments ([s] and [a] sounds) that are possibly confusing for listeners, may have hampered accent identification. In contrast, voicing temporal information may have facilitated listeners' task in 1-bit requantized as well as in lowpass-filtered speech.

**Amplitude envelope temporal information** Noise vocoded speech contains amplitude envelope temporal information, together with varying degrees of frequency domain information. 6-band noise vocoded speech allowed for accent identification above chance, whereas 3-band noise vocoded speech, containing very little frequency domain information, did not. Furthermore, when listeners could not process the remaining frequency domain information of 6-band noise vocoded speech because they were not given sentence transcripts of the (unintelligible) acoustic stimuli, accent identification performance was at chance. The lowpass-filtered stimuli used in the experiments also allowed for accent identification. They only contained information below 300 Hz, which primarily exhibited amplitude envelope temporal information. However, they may still have contained frequency domain information to some extent. Along with the results on noise vocoded stimuli, this suggests that listeners needed frequency domain information to complement amplitude envelope temporal information in order to identify speaker origin. The present experiments therefore suggest that amplitude envelope temporal information on its own is not particularly salient in terms of foreign accent.

## Frequency domain cues

For all the signal conditions that contained a certain amount of frequency domain information, we ran informal experiments where we asked listeners to discriminate vowels or consonants, given two signal-manipulated segments and two categories as options. For 1-bit requantized speech, [a] and [e] as well as [s] and [ʃ], but no other pair of vowels or fricatives, could be discriminated. However, when given two categories as options, the manner of articulation of consonants could usually be told apart. 6-band noise vocoded vowels as well as sibilants could be discriminated to some degree, with categories given as options. For 3-band noise vocoded or lowpass-filtered segments, this discrimination task

was no longer possible. In all signal conditions, it was typically not possible for listeners to identify single phones — they could only complete the task when given two categories as options. We conclude that the manner of articulation of consonants was to some degree accessible for listeners in 1-bit requantized speech; the same holds for vowel quality and the place of articulation of sibilants in 6-band noise vocoded (and therefore also *freqOnly*) speech.

**Spectral cues in general** Findings from 1-bit requantized, noise vocoded, and *sasasa*-speech presented in Chapter 4 suggest that the reduction of frequency domain information entails a decrease in accent identification performance. Subsequent experiments with lowpass-filtered and duration-transplanted speech were in line with this result: The presence of fundamental frequency contours in lowpass-filtered as opposed to monotonized lowpass-filtered speech facilitated accent identification, as did the exclusively spectral cues in *freqOnly* speech as opposed to the exclusively temporal cues in *timeOnly* speech. This last result in particular emphasizes that frequency domain information is highly salient: The strongly degraded spectral information in *freqOnly* speech yielded higher performance than the relatively detailed segmental (and therefore also suprasegmental) temporal information in *timeOnly* speech. Even though the monotonized lowpass-filtered stimuli were constructed so as to contain temporal information alone, certain cues to the articulation of consonants, for example, may have remained in the signal. If this were the case, it might explain the higher identification performance in this condition as opposed to *timeOnly* speech — together with the fact that different types of temporal information are conveyed by these signal conditions.

**Articulation of /r/** As reported in Chapter 6, the realization of /r/ biased results in the *timeOnly* signal condition: Listeners were inclined to perceive French-accented German when uvular /r/s were present in stimuli, and they tended to hear English-accented German when stimuli contained vocalized /r/s or no /r/. This, again, emphasizes the power of spectral information, and cues to the articulation of /r/ in particular, for the identification of speaker origin in a situation where the native languages tested significantly differ in their realization of this segment (Flege, 1984; Cunningham-Andersson and Engstrand, 1989; Vieru et al., 2011).

## Combined presence of time domain and frequency domain cues

We hypothesized in Part II of this thesis that temporal features may be of particular use when frequency domain information is strongly degraded, as typically occurs in listening situations that involve background noise or a telephone filter. The idea behind this was that some strongly degraded frequency domain cue may be of use to listeners because it occurs at a specific and expected moment in the time domain, and that it may be of no use to the listener if this time domain information is absent. We did in fact observe an additive trend when time domain and frequency domain information were combined, in 6-band noise vocoded speech, as opposed to both cues on their own, in *freqOnly* or *timeOnly* speech. However, both cues on their own also allowed for accent identification

above chance; the additivity can therefore be said to facilitate identification, but it is not necessary. In general, Table 9.1 illustrates that accent identification performance increases with increasing information available to the listener; this was the case when fundamental frequency contours were present in lowpass-filtered as opposed to monotonized lowpass-filtered speech, or when the amount of frequency domain information was increased in 6-band noise vocoded speech as opposed to 3-band noise vocoded speech. Similarly, 6-band noise vocoded speech without the presentation of sentence transcripts was assumed to hamper the processing of time domain and frequency domain information, as the speech signal was unintelligible in this condition.

## Acoustic measures of temporal information

In Chapter 3 we reported the results of 16 time domain measurements carried out on the French- and English-accented as well as on the native German material used in our perception experiments. The measures of articulation rate and pausing behavior did not reveal significant differences between the two non-native accents, nor did the applied rhythm metrics. The largest descriptive difference between French- and English-accented German was found in the percentage over which speech is voiced ($\%VO$). Further descriptive differences were revealed by the rate-normalized durational variability of adjacent voiced intervals ($nPVI\_VO$), the percentage over which speech is vocalic ($\%V$), and the rate-normalized durational variability of consonantal intervals ($varcoC$). These measures can thus be said to reflect speaker origin to some extent in our data, which is in line with some of the research discussed in Section 2.1.2. Quite possibly, the voicing temporal features facilitated listeners' identification performance in lowpass-filtered speech, together with amplitude envelope temporal information and the frequency domain information remaining below 300 Hz. However, accent identification performance based on temporal features alone was above chance for only one signal condition, *timeOnly*; for this signal condition, none of the 16 acoustic temporal measures was correlated with listeners' identification performance. These measures may reflect other features of non-native speech, such as speaker-individual (see Section 9.2) or general (see Section 9.3) properties of non-native speech. We therefore assume that acoustic correlates of listeners' identification performance in the signal condition *timeOnly* must be found in other characteristics of the time domain.

## Effect of accent

We found that French-accented German was recognized with higher accuracy than English-accented German in natural, 1-bit requantized, monotonized lowpass filtered, *freqOnly*, and *timeOnly* speech, but not in 6-band noise vocoded or lowpass-filtered speech with the original fundamental frequency contours. Certain features of French-accented German therefore seem to be more salient than those of English-accented German in the time domain as well as in the frequency domain. For *freqOnly* speech and possibly for 1-bit requantized speech, the French pronunciation of /r/ may have been highly salient to listeners. This would be in line with the /r/-driven bias discussed above. For the time

domain, these results suggest that in the absence of sufficient spectral characteristics, cues to voicing (absent in noise vocoded speech) carry features typical of a French accent. However, there must be an additional explanation: accent-specific cues to voicing were largely absent in *timeOnly* speech, yet French-accented German was recognized with higher accuracy in these signals. Furthermore, cues to voicing were present in lowpass-filtered as well as in monotonized lowpass-filtered speech, but only the latter yielded higher recognition performance for the French accent. Therefore, when the information available to listeners lies (almost) exclusively in the time domain, they perceive French-accented speech as more salient than English-accented speech. When more cues are available, e.g., fundamental frequency contours in lowpass-filtered speech, the salience of English-accented German seems to increase. This is in line with ideas brought forward in the context of research on speech rhythm: German and English — and, apparently, English-accented German — seem to be perceptually more similar in their rhythmic organization, while differing from French — and, apparently, French-accented German — in this regard (Lloyd James, 1929; Pike, 1945; Abercrombie, 1967; Ramus et al., 1999; Grabe and Low, 2002; Dellwo et al., 2007). Furthermore, this suggests that language-specific features of rhythmic organization are carried over to non-native speech. Evidence for this idea is reported in research by Ordin and Polyanskaya (2015), who find German learners of English to be more successful in acquiring target-like patterns of durational variability than French learners of English.

## Effect of speaker

We reported in Chapters 5 and 6 that listeners' accent identification performance varied according to speaker. Furthermore, this effect was not correlated with speakers' accent strength in *timeOnly* speech and only moderately correlated with accent strength in (monotonized) lowpass-filtered speech. We therefore assume that the nature of the influence of speaker origin on time domain characteristics differs between speakers. The idea that speakers employ different strategies in the time domain of their non-native speech is discussed in Section 9.2.

## Limitations of the present research

A number of limitations arose within the experiments probing the influence of speaker origin on temporal features of non-native speech. First, we observe that the search for signal manipulations that present (primarily) time domain characteristics often leads to signal types that do not occur in everyday communicative situations, which may hamper listeners' perception of foreign accent. But when applying the duration transplantation method in order to obtain maximally natural-sounding signals, the native German segmental (and, possibly, fundamental frequency contour) information produced certain artifacts in our results: listeners apparently used this segmental material as a cue in a task that was designed to be completed based on temporal information alone. Second, we were able to show that listeners could, in fact, identify the foreign accents in question based on temporal information alone, but listeners' performance was relatively poor. However, values obtained for listeners' sensitivity were comparable to those reported in experiments

on dialect (White et al., 2012) or language (Ramus et al., 2003) identification based on durational features. Third, it should be noted that the relative salience of different foreign accent features highly depends on the specific native language vs. non-native language combination in question (Cunningham-Andersson and Engstrand, 1989; Holm, 2008), and that identification performance is influenced by listeners' familiarity with speakers' native and non-native languages (Derwing and Munro, 1997; Pinet and Iverson, 2010). A wider range of native vs. non-native language combinations would therefore need to be subjected to the present experimental setup to be able to generalize the present results.

## 9.2 Evidence for speaker-individuality in the time domain

In the speech production experiment presented in Chapter 7, as well as in the perception experiment reported in Chapter 8, we discussed findings on speaker-individual characteristics that remain relatively invariant when speakers talk in different languages. For the production experiment, 16 speakers were recorded in their native Zurich German as well as in their non-native French and English. Temporal measures of pausing were applied to the materials and compared between speakers as well as within each speaker (i.e., across three languages). For the perception experiment, a subset of French and English sentences from each speaker was presented to native French and English listeners, respectively, who rated each sentence for accent strength on a continuous scale. This measure, i.e., accent strength, was also compared between speakers and within each speaker (i.e., across two languages).

We have shown in Chapters 7 and 8 that speakers' pausing behavior, namely the number of pauses and pause durations, as well as their accent strength is highly speaker-individual on the one hand, and on the other hand relatively robust towards within-speaker variability, even when speakers produced speech in different languages.

### Effect of language

The Zurich German speakers studied for this work produced fewer and shorter pauses in their native than in their non-native speech. This is intuitively sound: producing non-native speech is cognitively more demanding than producing native speech, and research has shown that cognitively more demanding tasks lead to an increase in pausing behavior (Grosjean, 1980; Riazantseva, 2001; de Jong et al., 2013). Furthermore, speakers produced fewer and shorter pauses in their non-native English than in their non-native French speech, and their foreign accent was perceived as being weaker in their English than in their French speech. This may point to the fact that the speakers studied are more proficient in English than in French. Despite the fact that French is an official language in Switzerland and that the speakers started learning French earlier than English at school, Swiss university students usually interact more often in English than in French.

## Effect of speaker

Speakers who produced many and long pauses in their native Zurich German speech most often behaved in the same way in their non-native French and English speech. This is in line with research by Derwing et al. (2009) and de Jong et al. (2013) who have found measures of speakers' non-native fluency to be influenced by speaker-individual fluency characteristics measured in their native language. It is further in line with research on speaker-specific temporal features across a range of within-speaker variation conditions, including the imitation of a different dialect (Dellwo et al., 2012, 2015; Leemann et al., 2014; Leemann and Kolly, 2015). This strongly suggests that speakers' pausing characteristics are determined not only by second language proficiency, but also to some extent by speaker-individual characteristics regarding cognitive factors (de Jong et al., 2013), reading ability or style (Laan, 1997), or anatomical characteristics such as lung volume that may influence breathing patterns (Dellwo et al., 2012, 2015).

Similarly, our Zurich German speakers' foreign accent strength as perceived by native listeners was highly variable between speakers, but rather invariant within a speaker: speakers perceived as having a strong accent in non-native French were usually perceived in a (proportionally) similar way in their non-native English. The extent to which one has a foreign accent therefore seems to be highly idiosyncratic unless a speaker aspires to moderate the strength of this accent (Moyer, 2004). Indeed, research on foreign accent strength has suggested that cognitive and social-psychological factors such as language attitudes and subjective identity influence accent strength (Guiora et al., 1972; Moyer, 2004; Kolly, 2011). Our further analyses showed that both our pausing measures were strongly correlated with strength of perceived foreign accent in speakers' French speech. This suggests that pausing is a strong cue for listeners' perception of a speakers' accent strength and complements existing research that investigated the importance of temporal characteristics such as segment and syllable durations on perceived accent strength (Tajima et al., 1997; Quené and van Delft, 2010; Winters and O'Brien, 2013). However, no such correlation arose in speakers' non-native English. This may have to do with the small number of pauses speakers produced in their non-native English; the English listeners may therefore have directed their attention to other characteristics of foreign-accented speech when rating accent strength. Flege (1988), for example, also found that pauses did not affect listeners' accent ratings.

## Effect of sentence

We observed an effect of sentence on measures of pausing; speakers tended to produce more and longer pauses in longer, syntactically complex sentences than in shorter and syntactically less complex sentences. This is intuitively sound, as the longer and syntactically more complex sentences present more slots for potential pauses. This is further in line with a wide range of time domain measures that have been shown to be affected by the sentence material used (Dellwo, 2010; Wiget et al., 2010; Arvaniti, 2012). There was also an effect of sentence on perceived accent strength in our data, where sentences

that contained a high number of marked and therefore difficult target features (e.g., nasal vowels in French) were rated particularly high in accent strength.

## 9.3 General properties of non-native speech in the time domain

A number of by-products from our work on the influence of speaker origin and speaker-individuality on non-native temporal features point to general properties of non-native speech. The Non-native Speaker Origin Corpus contained Standard German speech, spoken by Zurich German, French, and English speakers, whereas the Non-native Speaker-Individuality Corpus contained Zurich German, French, and English speech spoken by Zurich German speakers. Therefore, the differences we report between native speech and non-native speech are always based at least on two different non-native accents.

We found that non-native speech differs from native speech in a number of regards: First, non-native speech exhibited significantly lower articulation rates than native speech in the Non-native Speaker Origin Corpus; we thereby reproduced findings by, e.g., Trofimovich and Baker (2006) (see Chapter 3). Second, we found a higher number of silent pauses and longer pauses in non-native speech in both our corpora (see Chapters 3 and 7). This confirms research by Trofimovich and Baker (2006), Derwing et al. (2009) and de Jong et al. (2013): non-native speech is generally less fluent than native speech. Third, we found a number of rhythm metrics to vary between native and non-native speech in the Non-native Speaker Origin Corpus: *nPVI_V*, *varcoVOln*, *varcoUV*, and *nPVI_UV*. Non-native speech revealed significantly lower interval duration variability than native speech in the vocalic and voiced measures, while it exhibited significantly higher variability in the unvoiced measures. As for variability in vocalic and voiced measures, the finding may be explained by non-native speakers not (yet) mastering German vowel reduction (Dauer, 1983). Indeed, qualitative analyses of several examples revealed that non-native speakers tended to realize a high proportion of full vowels, whereas the native speakers usually reduced vowels in unstressed syllables. Furthermore, the native German speakers often elided unstressed vowels before liquids, therefore reducing not only the duration of certain vocalic, but also of certain voiced intervals. The finding that non-native speakers produce less durational variability in vocalic intervals corroborates results by Ordin and Polyanskaya (2015). The higher durational variability in non-native speakers' unvoiced intervals may be due to phenomena such as their tendency to pronounce an epenthetic velar plosive after velar nasals, thereby adding a short unvoiced segment and modifying the phonotactic structure of the target language, which may increase the durational variability of unvoiced intervals (see Chapter 6, Section 2.3).

## 9.4 Possible implications of the present work

Possible implications of the work conducted for the present thesis are twofold: results such as the ones presented here may be interesting for the field of forensic phonetics, and the findings may have implications in the domain of second language acquisition.

### Forensic phonetics

Incriminating speech samples for forensic voice comparison or forensic speaker profiling are frequently recorded over a telephone (Hirson et al., 1995), which degrades the spectral characteristics of the acoustic signal but does not affect temporal characteristics to the same extent (Chen et al., 2005). Similarly, speech samples for linguistic analysis for the determination of geographical origin in asylum cases are often recorded over a landline telephone (Baltisberger and Hubbuch, 2010).

Investigating listeners' accent identification performance in speech that contains only time domain cues and probing the additive effects of time domain and frequency domain cues in foreign accent identification may be beneficial for better analysis of speech recorded over a telephone connection or with background noise. Furthermore, time domain measures may complement a wide range of features from the frequency domain that point to speaker origin or speaker-individuality. In typical cases of forensic voice comparison, trace material from a crime such as recordings of a perpetrator during a bomb threat is compared to acoustic comparison material such as recordings of a suspect during a police interview. In such cases, it is desirable to apply acoustic measures that vary between speakers but are invariant within speakers, i.e., speaker-specific measures (Nolan, 2009). However, the Code of Practice of the International Association for Forensic Phonetics and Acoustics (2004) advises members to "exercise particular caution" when carrying out analyses on recordings of non-native speech. In fact, the impact of speaking a second language on speaker-individual characteristics is largely unknown, even though cases occur in forensic voice comparison where there is a mismatch in language between acoustic trace and comparison material (Herbert R. Masthoff, personal communication). More extensive research on between- and within-speaker variability regarding native and non-native speech is therefore desirable. The temporal pausing characteristics examined in the present work seem to meet the requirements for speaker-specific measures. Furthermore, the finding that accent strength is speaker-specific could also be leveraged for forensic cases where a speaker uses different non-native languages in different contexts, possibly in the presence of earwitnesses who may recall the strength of the speakers' accent. More research is needed before such results can be applied in forensic casework, however.

### Second language acquisition

As having a foreign accent can have consequences in communicative situations, some non-native speakers may wish to sound more native-like. This may be the case when their foreign accent affects the intelligibility of their speech (Derwing and Munro, 1997), but

also in cases where speakers are discriminated against because of their specific accent and origin (Lambert et al., 1960; Schairer, 1992; Cunningham-Andersson, 1996; Lippi-Green, 1997; Hirschfeld and Trouvain, 2007).

In cases where speakers wish to moderate their foreign accent, it is important to be aware of the acoustic characteristics that contribute to each particular accent. Our findings from perception experiments suggested that spectral characteristics and the pronunciation of /r/ are strong indicators of speaker origin. Regarding time domain characteristics, speakers' production of segment durations was relatively salient in terms of speaker origin. As shown by Tajima et al. (1997) and Quené and van Delft (2010), segment durations are also an important factor for listeners' ratings of speakers' intelligibility and accent strength. Furthermore, our findings from the production experiments suggest that temporal characteristics of voicing and, possibly, vocalic and consonantal intervals could be indicators of speaker origin. As for accent strength, pausing characteristics influence perceptual ratings of accent strength in Zurich German speakers' French, but not English, speech. A number of studies and teaching methods have dealt with the matter of teaching durational and rhythmic characteristics in the second language acquisition process (e.g., Wong, 1987; Fischer, 2007; Hirschfeld and Trouvain, 2007; Missaglia, 2007). However, as suprasegmental temporal features are known to be dependent on the intrinsic durations of the segments they contain (van Santen and Shih, 2000), it may be sufficient for non-native speakers to learn to produce segments of their non-native language, including their durations, in a target-like manner.

# 10

# Conclusions and outlook

This thesis comprises a number of speech perception and production experiments that have addressed temporal characteristics of non-native speech. In order to investigate the identification of speaker origin in non-native speech, a total of 130 subjects were tested in perception experiments using a wide range of signal manipulation methods that reduce frequency domain information to different degrees. This research was complemented by acoustic analyses of the temporal variability contained in the speech material used for stimulus creation. In order to investigate speaker-individual behavior cross-linguistically, a number of measures of temporal variability were applied to 16 speakers' native Zurich German and non-native French and English speech, and their accent strength, as perceived by native listeners, was investigated cross-linguistically. The collected data and subsequent analyses represent new approaches for the characterization of non-native speech in the time domain. In turn, this may have implications for the domains of forensic phonetics and second language acquisition.

## 10.1 Main contributions

**Cue weighting in foreign accent identification** Having developed an experimental setup and chosen a set of signal manipulation methods to present primarily or exclusively time domain information to listeners, we find that listeners can identify foreign accents based on time domain cues above chance. Segmental durations seem to be the most salient temporal cue in terms of foreign accent. Nevertheless, frequency domain cues are shown to be highly salient, and more so than time domain information. In particular, the pronunciation of /r/ is found to strongly influence listeners' perception of speaker origin in a situation where the languages tested significantly differ in their realization of this segment. We further find an additive trend, in that the combined presence of several cues strengthens listeners' identification performance.

**Evidence for accent-specific rhythmic features** We provide perceptual evidence that French-accented German sounds more salient in the time domain than English-accented German. This suggests that the rhythmic or temporal organization of French-accented German (and therefore of French) differs from German more strongly than that

of English-accented German (and therefore of English). This corroborates hypotheses from the domain of speech rhythm research, where the perceptual impression of French is argued to be more regular than that of English or German.

**Description of language-invariant speaker-individual features** Having constructed a database for studying between-speaker variability from a cross-linguistic point of view, we demonstrate that pausing characteristics as well as accent strength vary between speakers but remain relatively invariant within speakers when they produce speech in their native and two different non-native languages. This strongly suggests that pausing characteristics and accent strength are governed, to some extent, by speaker-individual cognitive, psychological, and anatomical factors.

## 10.2 Future work

Further systematic experimental investigations of the temporal cues that listeners use to identify speaker origin need to be developed and applied to a variety of native vs. non-native language combinations to better understand the influence of speaker origin on different non-native temporal features. Acoustic correlates of listeners' identification performance have to be sought by possibly developing new measures that characterize speech in the time domain. Furthermore, a wider range of acoustic measures needs to be tested to qualify and quantify how speaker-individual features influence the time domain of non-native speech in different languages.

**Development and application of further signal manipulation methods** We have tested as many as 10 different signal conditions that contain time domain and frequency domain information to various degrees. Some of them have a number of shortcomings (see Section 9.1). For example, to present listeners with monotonized lowpass-filtered speech that can be sure to contain no frequency domain information that would facilitate accent identification, the material could be filtered with a cutoff frequency of 180 Hz, as proposed by den Os (1988). This would convey amplitude envelope as well as voicing temporal information. A similar goal could be reached through the monotonization and subsequent delexicalization of speech material using the PURR-method (Prosody Unveiling Restricted Representation; Sonntag and Portele, 1998). To present only voicing temporal cues in a signal that may sound more natural than *sasasa*-speech, the amplitude of each period of lowpass-filtered or PURR signals could be set to a constant value. To present only amplitude envelope temporal information in a relatively natural-sounding signal, one could multiply the amplitude envelope information of noise vocoded speech with sinusoidal signals instead of white noise.

**Development of time domain measurements as correlates for accent identification performance** Our analyses do not reveal any acoustic measure that explains listeners' accent identification performance in a signal condition containing only time do-

main information. Other types of temporal measures may therefore need to be applied or developed in the search for acoustic correlates of perceptually salient temporal patterns.

**Application of the present experimental setup to other materials** We have presented results on listeners' accent identification performance for French- and English-accented German. It would now be sensible to apply the signal manipulation methods used and the experimental setup developed to other foreign accents. One could, for example, think of Italian- and Dutch-accented French, where similar assumptions could be made as to the temporal organization of these varieties: Italian-accented French may be closer to native French in its rhythmic and temporal organization than Dutch-accented French.

**Enlargement of the Non-native Speaker-Individuality Corpus** We have shown that multilingual speakers' temporal and rhythmic behavior is speaker-individual and language-independent to some extent. The experimental investigations carried out are based on 16 sentences per speaker and language. We know from research on speech rhythm that measures of temporal variability are strongly affected by the linguistic material contained in particular sentences, so it is desirable to analyze a larger number of sentences per speaker. A further step therefore involves the enlargement of this corpus, which has already been completed: it now contains 48 sentences per speaker and language.

**Application of further temporal measures to the Non-native Speaker-Individuality Corpus** Further steps in this research should include the application of a number of rhythm metrics to the Non-native Speaker-Individuality Corpus. We are currently carrying out an experiment that investigates between-speaker and within-speaker variability in automatically retrievable time domain measures (*%VO*, *varcoVOln*, *nPVI_Voiced*, *varcoPeak*, *nPVI_Peak*, see Chapter 3). These measures have been previously shown to vary between speakers and to be relatively robust against different types of within-speaker variation (see Section 2.1.3). We hope to discover further acoustic temporal measures that differ between speakers and are at the same time invariant within a speaker, i.e., between the different languages spoken by a speaker.

## 10.3 Concluding remarks

We argued in Chapter 1 that time domain characteristics of non-native speech could potentially differ from native speech in three different groups of features:

(i) features that are due to interference from a speaker's native language;

(ii) features that are influenced by speaker-individuality;

(iii) features that reflect general properties of non-native speech.

The speech perception and production experiments conducted for this thesis provide evidence for each of these groups of features. (i) Speaker origin manifested itself in a number

of temporal features; the proportion over which an utterance is voiced revealed the largest difference between French- and English-accented German speech production, whereas segment durations were shown to be the type of temporal cue most salient to listeners in terms of accent identification. (ii) Speaker-individuality appeared in two measures of pausing, namely the number and duration of pauses, and in perceived accent strength. Speaker-individuality remained apparent when participants produced speech in different languages. (iii) The description of general properties of non-native speech was not specifically aimed at by the present thesis; nevertheless, a number of such features became evident from the experiments conducted. Among these were lower general fluency, as measured by articulation rate and pausing behavior, as well as less durational variability in vocalic and voiced interval measures and more durational variability in unvoiced interval measures than exhibited in native speech.

Further investigations will be needed, however, to more fully understand the different factors determining non-native speech in the time domain.

# A

# Appendix: Reading materials

## Non-native Speaker Origin Corpus

1. Die Frau des Apothekers weiss immer, was sie will.
2. Das Theater hat viele neue Aufführungen geplant.
3. Er wollte sich seiner Schwächen einfach nicht bewusst werden.
4. Der öffentliche Verkehr lässt viel zu wünschen übrig.
5. Die schlechte Zahlungsbilanz lässt mich nicht zur Ruhe kommen.
6. Die Eltern geben ihm keine finanzielle Unterstützung.
7. Der starke Frühlingsregen hat grossen Schaden angerichtet.
8. Der schnellste Zug ist immer noch der ICE.
9. Der Wiederaufbau der Stadt wird sehr lange dauern.
10. Das Bildungsministerium hat den einfachsten Weg gewählt.
11. Diese Konditorei macht ausgezeichnete Kuchen.
12. Dieses Geschäft bietet sehr preisgünstige Ware an.
13. Sie haben die Wahrheit erst entdeckt, als er auspackte.
14. Für meine Mannschaft wird der Sieg ein Kinderspiel sein.
15. Die Meinungsumfragen sagen einen Sieg der Rechten voraus.
16. Die Strassen der Innenstadt wurden von der Polizei gesperrt.
17. Ein berühmtes Bild wurde aus dem Kunsthaus gestohlen.
18. Der Müssiggang ist bekanntlich aller Laster Anfang.

# Non-native Speaker-Individuality Corpus (subset of TEVOID Corpus)

## Zurich German sentences

1. Chasch ja nöd nöime andersch go studiere mit Erasmus.
2. Mäischtens ladt mich min Papi ii.
3. Uf jede Fall händ's s Gfühl, äine hät en Alarm truckt.
4. Ich glaub vo de Temperatur här isch es nöd würkli chelter gsi.
5. Politik hät mich au früener scho sehr intressiert.
6. Es Semeschter vorher han i über Südoschtasie es Seminar gmacht.
7. De ganz Überfall wird äigentlich dadurch gschtöört, dass öpper versuecht bi de Iigangstüür ine z choo.
8. Japan bruucht ja vor alem im Summer vil Elektrizität.
9. Irgendwie hät no äine wele i d Bank ine.
10. Ich bi würkli a allem interessiert.
11. Maskierti Persone händ sich Zuetritt i de Bank verschafft.
12. Z ersch häsch so gseh wie si de Tresor uufgmacht händ.
13. Dur das hät de Trainer halt volli Kontrolle über öises Läbe ghaa.
14. Nachethäär han i so chli zu palestinänsischem Terrorismus gschaffet.
15. Ich han scho vorher im Film gschafft im e andere Beräich.
16. Äntwäder hät mer ghulfe serviere döte, im Huus, oder mer isch useggange.

## French sentences

1. Mais tu ne peux pas aller étudier ailleurs avec Erasmus.
2. La plupart du temps c'est mon Papa qui m'invite.
3. En tout cas ils ont l'impression que quelqu'un a appuyé sur l'alarme.
4. Je crois qu'au niveau de la température il ne faisait pas vraiment plus froid.
5. La politique m'a déjà beaucoup intéressée par le passé.

6. Dans le semestre précédent j'ai fait un séminaire sur l'Asie du sud-est.
7. En fait, tout le cambriolage est dérangé parce que quelqu'un essaye d'entrer par la porte d'entrée.
8. Le Japon a besoin de beaucoup d'électricité surtout en été.
9. D'une manière ou d'une autre, quelqu'un a encore voulu entrer dans la banque.
10. Je suis vraiment intéressée à tout.
11. Des personnes masquées se sont introduites dans cette banque.
12. D'abord tu as vu comment ils ont ouvert le coffre-fort.
13. Ainsi, l'entraîneur avait un contrôle total sur notre vie.
14. Après ça j'ai fait un peu de travail sur le terrorisme palestinien.
15. Déjà avant j'ai travaillé dans le film, dans un autre domaine.
16. Soit on aidait à servir là-bas, dans la maison, soit on allait dehors.

## English sentences

1. Well, you cannot go studying somewhere else with Erasmus.
2. Usually my dad invites me.
3. In any case, they have the feeling that someone pressed the alarm.
4. I believe that in terms of temperature it was not really colder.
5. Politics has already interested me back in the days.
6. In the previous semester I attended a seminar on South-East Asia.
7. The entire robbery is disturbed by someone trying to enter the front door.
8. Japan needs a lot of electricity, especially in the summer.
9. Somehow, someone also wanted to enter the bank.
10. I am really interested in everything.
11. Masked people gained access to this bank.
12. First you saw how they opened the safe.
13. In this way, the trainer had total control over our lives.
14. Afterwards I did a little work on Palestinian terrorism.

15. I have worked with films before in a different area.
16. One could either help serving in the house, or go outside.

# B

# References


Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press, Edinburgh.

Adams, C. and Munro, R. R. (1978). In search of acoustic correlates of stress: Fundamental frequency, amplitude and duration in the connected utterance of some native and non-native speakers of English. *Phonetica*, 35:125–156.

Allen, J. S., Miller, J. L., and DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America*, 113:544–552.

Arslan, L. M. and Hansen, J. H. L. (1997). A study of temporal features and frequency characteristics in American English foreign accent. *Journal of the Acoustical Society of America*, 102(1):28–40.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40:351–371.

Asu, E. L. and Nolan, F. (2006). Estonian and English rhythm: A two-dimensional quantification based on syllables and feet. *Phonetica*, 66:64–77.

Baltisberger, E. and Hubbuch, P. (2010). LADO with specialized linguists – the development of LINGUA's working method. In Zwaan, K., Verrips, M., and Muysken, P., editors, *The role of language in European asylum procedures*, pages 9–19. Wolf Legal Publishers, Nijmegen.

Barry, W., Andreeva, B., and Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica*, 66:1–17.

Bates, D. M. and Maechler, M. (2009). *lme4: Linear mixed-effects models using S4 classes. R package version 1.1-7*.

Bent, T., Bradlow, A. R., and Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica*, 65:131–147.

Boersma, P. and Weenink, D. (2012). *Praat. Doing phonetics by computer.* http://www.praat.org.

Braun, A. and Rosin, A. (2015). On the speaker-specificity of hesitation markers. In *Proceedings of the International Congress of Phonetic Sciences 2015*, pages 1–4, Glasgow.

Caramazza, A., Yeni-Komshian, G., Zurif, E., and Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America*, 54(2):421–428.

Carter, P. M. (2005). Quantifying rhythmic differences between Spanish, English, and Hispanic English. In Randall, S. G. and Edward, J. R., editors, *Theoretical and experimental approaches to romance linguistics*, pages 63–75. Benjamins, Amsterdam.

Chen, B., Zhu, Q., and Morgan, N. (2005). Long-term temporal features for conversational speech recognition. In Bengio, S. and Bourlard, H., editors, *Machine learning for multimodal interaction*, pages 232–242. Springer, Berlin/Heidelberg/New York.

Classe, A. (1939). *The rhythm of English prose.* Blackwell, Oxford.

Coetzee, A. W., García-Amaya, L., Henriksen, N., and Wissing, D. (2015). Bilingual speech rhythm: Spanish-Afrikaans in Patagonia. In *Proceedings of the International Congress of Phonetic Sciences 2015*, pages 1–4, Glasgow.

Cumming, R. E. (2011). Perceptually informed quantification of speech rhythm in pairwise variability indices. *Phonetica*, 68:256–277.

Cummins, F. and Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, 26:145–171.

Cunningham-Andersson, U. (1996). Native speaker reactions to non-native speech. In James, A. and Leather, J., editors, *Second language speech. Structure and process*, pages 133–144. Mouton de Gruyter, Berlin/New York.

Cunningham-Andersson, U. and Engstrand, O. (1989). Perceived strength and identity of foreign accent in Swedish. *Phonetica*, 46:138–154.

Dasher, R. and Bollinger, D. (1982). On pre-accentual lengthening. *Journal of the International Phonetic Association*, 12:58–69.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11:51–62.

de Jong, N. H., Groenhout, R., Schoonen, R., and Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 34:1–21.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In Karnowski, P. and Szigeti, I., editors, *Language and language-processing*, pages 231–241. Lang, Frankfurt am Main.

Dellwo, V. (2008). The role of speech rate in perceiving speech rhythm. In *Proceedings of Speech Prosody 2008*, pages 375–378, Campinas.

Dellwo, V. (2010). *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence*. PhD thesis, University of Bonn, Bonn.

Dellwo, V., Fourcin, A., and Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In *Proceedings of the International Congress of Phonetic Sciences 2007*, pages 1129–1132, Saarbrücken.

Dellwo, V., Leemann, A., and Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *Proceedings of Interspeech 2012*, pages 1584–1587, Portland, OR.

Dellwo, V., Leemann, A., and Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *Journal of the Acoustical Society of America*, 137(3):1513–1528.

Dellwo, V. and Schmid, S. (2015). Speaker-individual rhythmic characteristics in read speech of German-Italian bilinguals. In Leemann, A., Kolly, M.-J., Schmid, S., and Dellwo, V., editors, *Trends in phonetics and phonology: Studies from German-speaking Europe*, pages 349–362. Lang, Bern.

den Os, E. (1988). *Rhythm and tempo in Dutch and Italian*. Elinkwijk, Utrecht.

Derwing, T. M. and Munro, M. J. (1997). Accent, intelligibility and comprehensibility. Evidence from four L1's. *Studies in Second Language Acquisition*, 20:1–16.

Derwing, T. M., Munro, M. J., Thomson, R. I., and Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, 31:533–557.

Dorn, A., O'Reilly, M., and Ní Chasaide, A. (2012). A preliminary analysis of speech rhythm in three varieties of Irish (Gaelic). In *Proceedings of Speech Prosody 2012*, pages 510–513, Shanghai.

Ellis, S. (1994). The Yorkshire Ripper enquiry: Part 1. *Forensic Linguistics*, 1:197–206.

Fant, G. (1960). *Acoustic theory of speech production*. Mouton & Co, The Hague.

Ferragne, E. and Pellegrino, F. (2004). Rhythm in read British English: Interdialect variability. In *Proceedings of the International Conference on Spoken Language Processing 2004*, pages 1573–1576, Jeju.

Fischer, A. (2007). *Deutsch lernen mit Rhythmus. Der Sprechrhythmus als Basis einer integrierten Phonetik im Unterricht Deutsch als Fremdsprache*. Schubert, Leipzig.

Flege, J. and Hillenbrand, J. (1984). Limits on phonetic accuracy in foreign language speech production. *Journal of the Acoustical Society of America*, 76(3):708–721.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, 76(3):692–707.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in English sentences. *Journal of the Acoustical Society of America*, 84(1):70–79.

Flege, J. E. (1992). Speech learning in a second language. In Ferguson, C. A., Menn, L., and Stoel-Gammon, C., editors, *Phonological development. Models, research, implications*, pages 565–604. York Press, Maryland.

Flege, J. E., Munro, M. J., and Skelton, L. (1992). Production of the word-final English /t/-/d/ contrast by native speakers of English, Mandarin, and Spanish. *Journal of the Acoustical Society of America*, 92(1):128–143.

Flege, J. E. and Port, R. (1981). Cross-language phonetic interference: Arabic to English. *Language and Speech*, 24(2):125–146.

Fowler, C. A., Sramko, V., Ostry, D. J., Rowland, S. A., and Hallé, P. (2008). Cross language phonetic influences on the speech of French-English bilinguals. *Journal of Phonetics*, 36(4):649–663.

Goldman Eisler, F. (1968). *Psycholinguistics. Experiments in spontaneous speech.* Academic Press, London/New York.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In Gussenhoven, C. and Warner, N., editors, *Laboratory Phonology 7*, pages 515–545. Mouton de Gruyter, Berlin/New York.

Grenon, I. and White, L. (2008). Acquiring rhythm. A comparison of L1 and L2 speakers of Canadian English and Japanese. In *Proceedings of the Boston University Conference on Language Development 2008*, pages 155–166, Boston.

Grosjean, F. (1980). Temporal variables within and between speakers. In Dechert, H. W. and Raupach, M., editors, *Towards a cross-linguistic assessment of speech production*, pages 39–53. Lang, Frankfurt am Main.

Guiora, A., Beit-Hallahmi, B., Brannon, R. C., and Scovel, T. (1972). The effects of experimentally induced changes in ego states on pronunciation ability in a second language: An exploratory study. *Comprehensive Psychiatry*, 13(5):421–428.

Gutiérrez Díez, F., Dellwo, V., Gavaldà, N., and Rosen, S. (2008). The development of measurable speech rhythm during second language acquisition. *Journal of the Acoustical Society of America*, 123(5):3886.

Hazan, V. L. and Boulakia, G. (1993). Perception and production of a voicing contrast by French-English bilinguals. *Language and Speech*, 36(1):17–38.

Hirschfeld, U. and Trouvain, J. (2007). Teaching prosody in German as a foreign language. In Trouvain, J. and Gut, U., editors, *Non-native prosody. Phonetic description and teaching practice*, pages 171–187. Mouton de Gruyter, Berlin/New York.

Hirson, A., French, P., and Howard, D. (1995). Speech fundamental frequency over the telephone and face-to-face: Some implications for forensic phonetics. In Windsor Lewis, J., editor, *Studies in general and English phonetics in honour of Professor J. D. O'Connor.*, pages 230–240. Routledge, London.

Holm, S. (2008). *Intonational and durational contributions to the perception of foreign-accented Norwegian. An experimental phonetic investigation.* PhD thesis, Norwegian University of Science and Technology, Trondheim.

International Association for Forensic Phonetics and Acoustics (2004). *IAFPA Code of Practice.* http://iafpa.net/code.htm, accessed 18.12.2015.

Kliegl, R., Wei, P., Dambacher, M., Yan, M., and Zhou, X. (2011). Experimental effects and individual differences in linear mixed models: Estimating the relationship between spatial, object, and attraction effects in visual attention. *Frontiers in Psychology*, 1:1–12.

Kohler, K. J. (2009). Rhythm in speech and language. A new research paradigm. *Phonetica*, 66(1-2):29–45.

Kolly, M.-J. (2011). Weshalb hat man (noch) einen Akzent? Eine Untersuchung im Schnittfeld von Akzent und Einstellung bei Schweizer Dialektsprechern. *Linguistik online*, 50(6):43–77.

Köster, O., Kehrein, R., Masthoff, K., and Boubaker, Y. H. (2012). The tell-tale accent: Identification of regionally-marked speech in German telephone conversations by forensic phoneticians. *Journal of Speech, Language and the Law*, 19(1):51–71.

Künzel, H. J. (2013). Some general phonetic and forensic aspects of speaking tempo. *Journal of Speech, Language and the Law*, 4(1):48–83.

Laan, G. (1997). The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22:43–65.

Lambert, W. E., Hodgson, R. C., Gardner, R. C., and Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology*, 80(1):44–51.

Lee, C. S. and Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition*, 93(3):225–254.

Leemann, A., Dellwo, V., Kolly, M.-J., and Schmid, S. (2012). Rhythmic variability in Swiss German dialects. In *Proceedings of Speech Prosody 2012*, pages 607–610, Shanghai.

Leemann, A. and Kolly, M.-J. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication*, 75:97–122.

Leemann, A., Kolly, M.-J., and Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, 238:59–67.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5:253–263.

Lippi-Green, R. (1997). *English with an accent: Language, ideology and discrimination in the United States*. Routledge, London/New York.

Lloyd James, A. (1929). *Historical introduction to French phonetics*. University of London Press, London.

Loukina, A., Kochanski, G., Rosner, B., and Keane, E. (2011). Rhythm measures and dimensions of durational variation in speech. *Journal of the Acoustical Society of America*, 129(5):3258–3270.

Loula, F., Prasad, S., Harber, K., and Shiffrar, M. (2005). Recognizing people from their movement. *Journal of Experimental Psychology: Human Perception and Performance*, 31(1):210–220.

Low, E. L. (1998). *Prosodic prominence in Singapore English*. PhD thesis, University of Cambridge, Cambridge.

Low, E. L., Grabe, E., and Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, 43:377–401.

Major, R. C. (2001). *Foreign accent. The ontogeny and phylogeny of second language phonology*. Erlbaum, Mahwah NJ/London.

McAllister, R., Flege, J., and Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics*, 30:229–258.

McDougall, K. (2004). Speaker-specific formant dynamics: An experiment on Australian English /aI/. *Journal of Speech, Language and the Law*, 11(1):103–130.

McDougall, K. (2006). Dynamic features of speech and the characterization of speakers: Towards a new approach using formant frequencies. *Journal of Speech, Language and the Law*, 13(1):89–126.

Mermelstein, P. (1975). Automatic segmentation of speech into syllabic units. *Journal of the Acoustical Society of America*, 58(4):626–637.

Missaglia, F. (2007). Prosodic training for adult Italian learners of German: The Contrastive Prosody method. In Trouvain, J. and Gut, U., editors, *Non-native prosody. Phonetic description and teaching practice*, pages 237–258. Mouton de Gruyter, Berlin/New York.

Moyer, A. (2004). *Age, accent and experience in second language acquisition. An intergrated approach to critical period inquiry.* Multilingual Matters, Buffalo NY.

Munro, M. J. and Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, 23(4):451–468.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3):756–766.

Nolan, F. (2009). *The phonetic bases of forensic speaker identification.* Cambridge University Press, Cambridge, 2nd edition.

Nolan, F. and Asu, E. L. (2009). The Pairwise Variability Index and coexisting rhythms in language. *Phonetica*, 66(1-2):64–77.

Nolan, F. and Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B*, 369:20130396.

Obin, N., Avanzi, M., Bordal, G., and Bardiaux, A. (2012). Regional variations of speech rhythm in French: In search of lost times. In *Proceedings of Speech Prosody 2012*, pages 406–409, Shanghai.

Ordin, M. and Polyanskaya, L. (2015). Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *Journal of the Acoustical Society of America*, 138:533–544.

Perrier, P. (2012). Gesture planning integrating knowledge of the motor plant's dynamic: A literature review from motor control and speech motor control. In Fuchs, S., Weirich, M., Pape, D., and Perrier, P., editors, *Speech planning and dynamics*, pages 191–238. Lang, Frankfurt am Main.

Pike, K. L. (1945). *The intonation of American English.* University of Michigan Press, Ann Arbor.

Pinet, M. and Iverson, P. (2010). Talker-listener accent interactions in speech-in-noise recognition: Effects of prosodic manipulation as a function of language experience. *Journal of the Acoustical Society of America*, 128(3):1357–1365.

Quené, H. and van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, 52:911–918.

R Core Team (2014). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, 2:85–115.

Ramus, F., Dupoux, E., and Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the International Congress of Phonetic Sciences 2003*, pages 337–342, Barcelona.

Ramus, F. and Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105(1):512–521.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292.

Riazantseva, A. (2001). Second language proficiency and pausing: A study of Russian speakers of English. *Studies in Second Language Acquisition*, 23:497–526.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In Crystal, D., editor, *Linguistic controversies*, pages 73–79. Arnold, London.

Schairer, K. E. (1992). Native speaker reaction to non-native speech. *Modern Language Journal*, 76(3):309–319.

Schmid, S. (2012). Phonological typology, rhythm types and the phonetics-phonology interface: A methodological overview and three case studies on Italo-Romance dialects. In Ender, A., Leemann, A., and Wälchli, B., editors, *Methods in contemporary linguistics. A Festschrift in honour of Iwar Werlen*, pages 45–68. Mouton de Gruyter, Berlin/New York.

Shriberg, E., Ferrer, L., Kajarekar, S., Venkataraman, A., and Stolcke, A. (2005). Modelling prosodic feature sequences for speaker recognition. *Speech Communication*, 46:455–472.

Sonntag, G. P. and Portele, T. (1998). A method for prosody evaluation and investigation. *Computer, Speech and Language*, 12(4):437–451.

Tajima, K., Port, R., and Dalby, J. (1997). Effects of temporal correction in intelligibility of foreign-accented English. *Journal of Phonetics*, 25:1–24.

Taylor, D. S. (1981). Nonnative speakers and the rhythm of English. *International Review of Applied Linguistics in Language Teaching*, 19:221–226.

Tilsen, S. and Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America*, 134(1):628–639.

Tilsen, S. and Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*, 124(2):EL34–EL39.

Tortel, A. and Hirst, D. (2010). Rhythm metrics and the production of English L1/L2. In *Proceedings of Speech Prosody 2010*, pages 1–4.

Trofimovich, P. and Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition*, 28:1–30.

van Santen, J. P. H. and Shih, C. (2000). Suprasegmental and segmental timing models in Mandarin Chinese and American English. *Journal of the Acoustical Society of America*, 107(2):1012–1026.

Vieru, B., Boula de Mareüil, P., and Adda-Decker, M. (2011). Characterisation and identification of non-native French accents. *Speech Communication*, 53(3):292–310.

White, L. and Mattys, S. L. (2007a). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35:501–522.

White, L. and Mattys, S. L. (2007b). Rhythmic typology and variation in first and second languages. In Prieto, P., Mascaró, J., and Solé, M.-J., editors, *Segmental and prosodic issues in Romance phonology*, pages 237–257. Benjamins, Amsterdam/Philadelphia.

White, L., Mattys, S. L., and Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language*, 66:665–679.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America*, 127(3):1559–1569.

Winters, S. and O'Brien, M. G. (2013). Perceived accentedness and intelligibility: The relative contributions of F0 and duration. *Speech Communication*, 55:486–507.

Wong, R. (1987). *Teaching pronunciation: Focus on English rhythm and intonation.* Prentice Hall Regents, Englewood Cliffs, NJ.

Yoon, T.-J. (2010). Capturing inter-speaker invariance using statistical measures of speech rhythm. In *Proceedings of Speech Prosody 2010*, pages 1–4, Chicago.